(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2023年6月29日(29.06.2023)

WIPO | PCT

(10) 国際公開番号
WO 2023/120403 A1

(51) 国際特許分類:
G06F 17/16 (2006.01) *G06F 17/10* (2006.01)
G06F 7/24 (2006.01)

(21) 国際出願番号： PCT/JP2022/046353

(22) 国際出願日： 2022年12月16日(16.12.2022)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(30) 優先権データ：
特願 2021-209979 2021年12月23日(23.12.2021) JP

(71) 出願人: 国立大学法人奈良先端科学技術大学院大学 (**NATIONAL UNIVERSITY CORPORATION NARA INSTITUTE OF SCIENCE AND TECHNOLOGY**) [JP/JP]; 〒6300192 奈良県生駒市高山町８９１６番地の５ Nara (JP).

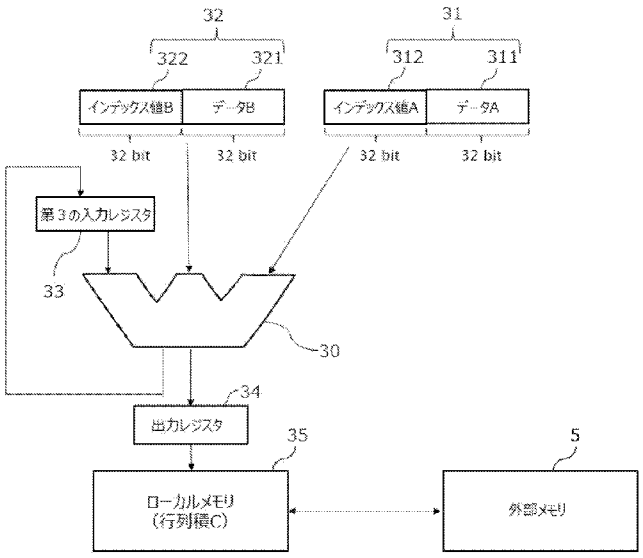
(72) 発明者: 中島 康彦 (**NAKASHIMA, Yasuhiko**); 〒6300192 奈良県生駒市高山町８９１６番地の５ 国立大学法人奈良先端科学技術大学院大学内 Nara (JP). 船井 遼太朗 (**FUNAI, Ryotaro**); 〒6300192 奈良県生駒市高山町８９１６番地の５ 国立大学法人奈良先端科学技術大学院大学内 Nara (JP).

(74) 代理人: 弁理士法人グローバル知財 (**GLOBAL INTELLECTUAL PROPERTY**); 〒6500021 兵庫県神戸市中央区三宮町３丁目７－６ 神戸元町ユニオンビル９Ｆ Hyogo (JP).

(54) **Title:** CALCULATION UNIT INVOLVED IN MERGING AND SORTING AND PERFORMING SPARSE MATRIX COMPUTATION BY CGRA

(54) 発明の名称： ＣＧＲＡによる疎行列計算とマージソートに関する演算ユニット

[図6]



5 External memory
33 Third input register
34 Output register
35 Local memory (matrix product C)
311 Data A
312 Index value A
321 Data B
322 Index value B

(57) **Abstract:** Provided is a unit that is for calculation by CGRA, that effectively uses a memory space, and that makes it possible to obtain a sparse matrix product by using less local memory. The calculation unit comprises a 3-input pipeline floating point sum-of-product calculator and a local memory. A sum-of-product calculator 30 is provided with: a first input register 31 and a second input register 32 that each are composed of a set of an index register and a data register; and a third input register 33 and an output register 34 that are data registers. In the first and second input registers of the

WO 2023/120403 A1

［続葉有］

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一　国際調査報告（条約第21条(3)）

sum-of-product calculator, index values A, B read from the local memory are stored in the respective index registers, data A, B read from the local memory are stored in the respective data registers, and a comparison is made between the index value A and the index value B. If the index values match with each other, a multiplication value of data A, B and a data value of the third input register are added, and the resultant value is stored in the output register of the sum-of-product calculator and also returns to the third input register.

(57) 要約：メモリ空間を有効に利用し、少ないローカルメモリを用いて疎行列積を可能とするＣＧＲＡによる演算ユニットを提供する。演算ユニットは、３入力パイプライン浮動小数点積和演算器とローカルメモリから構成され、積和演算器３０は、インデックスレジスタとデータレジスタを組にした第１の入力レジスタ３１及び第２の入力レジスタ３２、データレジスタである第３の入力レジスタ３３及び出力レジスタ３４を備える。積和演算器の第１及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａ，Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡ，Ｂが各データレジスタに格納され、インデックス値Ａとインデックス値Ｂが比較され、値が一致する場合には、データＡ，Ｂの乗算値と第３の入力レジスタのデータ値が加算されて積和演算器の出力レジスタに格納されると共に、第３の入力レジスタに戻る。

明　細　書

発明の名称 ：

ＣＧＲＡによる疎行列計算とマージソートに関する演算ユニット

技術分野

[0001] 本発明は、２次元アレイ状に配置されたＰＥ（Processing Element）を計算資源に持つＣＧＲＡ（Coarse-grained reconfigurable architecture）における演算ユニットに関し、特に、疎行列計算とマージソートに関する演算ユニットに関するものである。

背景技術

[0002] 昨今、ＩｏＴ（Internet of Things）など高い電力あたりの性能を持つデバイスが求められ、主にループ処理などをＣＰＵの代わりに実行するアクセラレータとして有用なＣＧＲＡが注目されている。

ＣＧＲＡは、シストリックアレイを起源とするアーキテクチャーで、ＦＰＧＡ（Field Programmable Gate Array）と同様に回路を何度でも変更できる再構成可能なアーキテクチャーである。ＣＧＲＡは演算ユニット及びレジスタ等からなる基本ユニットとメモリから構成されることが多く、再構成の粒度がＦＰＧＡではゲートレベルであるのに対し、ＣＧＲＡでは演算ユニットレベルである。そのため、ＣＧＲＡは、同等プロセスルールで同等機能を実現する場合には、ＦＰＧＡに比べ、動作周波数や回路面積で優位である。また、ＣＧＲＡは、パイプライン型垂直方向並列処理であり、演算ユニット群に毎サイクル供給するデータを抑えることが可能である。

[0003] ニューラルネットワークの計算などにおいて、大規模なパラメータ行例の計算をプロセッサで処理する際、計算量が多くなり処理負担が大きくなることから、処理負担を軽くすべく、主にＦＰＧＡに関する疎行列ベクトル積を並列計算するアクセラレータにおいて、必要とするハードウエア資源を削減するアクセラレータが知られている（特許文献１を参照）。

特許文献１のアクセラレータは、データロード部が記憶部に記憶している

入力ベクトルの要素の各々に対して並列にアクセスする順序に従い、疎行列の非零要素の各々の配置にパディングを挿入する。しかしながら、特許文献１のアクセラレータにおける疎行列ベクトル積は、データバッファからの入力要素と重みパラメータ列の融合積和を行うものであり、これは密行列－疎行列の行列積となり、疎行列－疎行列の行列積に関するものではない。

[0004] また、データセンター向けに、ＣＧＲＡを用いた密行列テンソル計算アクセラレータが実用化されているが、画像フィルタなど実際の畳み込みニューラルネットワーク（ＣＮＮ）の計算を行うＡＩ（Artificial Intelligence）アクセラレータでは、零要素の多い疎行列を取り扱うことが多い。しかしながら、大容量メモリを利用できるデータセンターでは、零要素を取り除くことなく実行しているのが実状である。

先行技術文献

特許文献

[0005] 特許文献1：特開２０２０－１６６３６８号公報

発明の概要

発明が解決しようとする課題

[0006] メモリ容量に制約がある超小型のＡＩアクセラレータでは、零要素の多い疎行列を取り扱う場合、零要素を削除して、限りあるメモリ空間を有効に利用できる疎行列－疎行列の積を行うべきである。

かかる状況に鑑みて、本発明は、メモリ空間を有効に利用し、少ないローカルメモリを用いて疎行列－疎行列の積とマージソートを可能とするＣＧＲＡによる疎行列計算とマージソートに関する演算ユニットを提供することを目的とする。

課題を解決するための手段

[0007] 上記課題を解決すべく、本発明の演算ユニットは、３入力パイプライン浮動小数点積和演算器とローカルメモリから構成される演算ユニットであって、積和演算器は、インデックスレジスタとデータレジスタを組にした第１の

入力レジスタ及び第２の入力レジスタ、データレジスタである第３の入力レジスタ及び出力レジスタを備える。積和演算器の第１の入力レジスタ及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納される。そして、インデックス値Ａとインデックス値Ｂが比較され、値が一致する場合には、データＡとデータＢの乗算値と第３の入力レジスタのデータ値が加算されて積和演算器の出力レジスタに格納されると共に、第３の入力レジスタに戻る。

[0008] 上記構成によれば、ＣＧＲＡ型アクセラレータに、データパスを追加して、少ないローカルメモリを用いた疎行列－疎行列の行列積の演算を可能とする。具体的には、行列の要素番号であるインデックス値を上位３２ｂｉｔ、行列の要素値であるデータを下位３２ｂｉｔ格納する６４ｂｉｔデータ（８ｂｙｔｅ）をローカルメモリに格納し、疎行列－疎行列の行列積の演算を可能とする。また、２つの入力が、疎行列であることに着目して、零要素を取り除き、データを圧縮して計算を行い、簡素な機構により、より高度な疎行列－疎行列の積を、疎行列－密行列を同じスループットで高効率に計算することを可能とする。ＣＧＲＡはパイプライン型垂直方向であるため、垂直方向（縦方向）の圧縮はできないことから、水平方向（横方向）の零要素を取り除きデータを圧縮する。そして、データパスを追加することにより、少ないローカルメモリを用いた疎行列計算を途切れることなく可能とする。

[0009] 本発明の演算ユニットは、具体的には、疎行列Ａと疎行列Ｂの行列積の演算において、インデックス値Ａは疎行列の列番号で、インデックス値Ｂは疎行列の転置行列の列番号であり、各疎行列におけるゼロ値の要素は圧縮され、非ゼロ値の要素がインデックス値とデータを組としてローカルメモリに記憶される。

ここで、疎行列におけるゼロ値の要素が圧縮されるとは、ゼロ値のものを取り除き、非ゼロ値のものだけを残して、列番号を前に詰めることである。

インデックス値には、元の列番号が格納され、データには元の列番号の値が格納されている。

[0010] 本発明の演算ユニットにおいて、クロック信号のタイミングｎ＋１で、積和演算器の第１段の第１の入力レジスタ及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納される。

そして、クロック信号のタイミングｎ＋２で、インデックス値Ａとインデックス値Ｂが比較され、値が異なる場合には定数Ｃが積和演算器の第１段の出力データレジスタに格納され、値が一致する場合にはデータＡとデータＢの乗算が実行されて積和演算器の第１段の出力データレジスタに格納される。

クロック信号のタイミングｎ＋３で、積和演算器の第２段の出力データレジスタに加算結果が伝搬され、クロック信号のタイミングｎ＋４で、積和演算器の第３段の出力データレジスタに正規化済の積和演算結果が伝搬される。

そして、クロック信号のタイミングｎ＋５で、積和演算器の第３段の出力データレジスタから積和演算器の第１段の第３の入力レジスタであるデータレジスタに戻される。

[0011] 本発明の演算ユニットは、入力ベースアドレスレジスタと入力オフセットアドレスレジスタを含む第１のアドレス計算機構および第２のアドレス計算機構を備え、第１および第２のアドレス計算機構は、各々の入力ベースアドレスレジスタおよび入力オフセットアドレスレジスタにアドレス情報をロードし、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂを比較し、第１のアドレス計算機構は、Ａ≦Ｂの場合に入力ベースアドレスレジスタに次要素参照に必要な値を加算し、第２のアドレス計算機構は、Ａ≧Ｂの場合に入力ベースアドレスレジスタに次要素参照に必要な値を加算して、アドレス加算結果を入力ベースアドレスレジスタに戻す。

[0012] 本発明の演算ユニットにおいて、クロック信号の第１のタイミングで、各々の入力ベースアドレスレジスタおよび入力オフセットアドレスレジスタにアドレス情報をロードし、クロック信号の第２のタイミングで、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂを比較し、第１のアドレス計算機構は、Ａ≦Ｂの場合に入力ベースアドレスレジスタに次要素参照に必要な値を加算し、第２のアドレス計算機構は、Ａ≧Ｂの場合に入力ベースアドレスレジスタに次要素参照に必要な値を加算し、クロック信号の第３および第４のタイミングで、アドレス加算結果をデータ保持レジスタに伝搬させ、クロック信号の次のタイミングで、データ保持レジスタの内容を入力ベースアドレスレジスタに戻す。

[0013] 本発明の演算ユニットにおいて、第１及び第２のアドレス計算機構は、ローカルメモリを複数空間に分割するアドレスマスク機構を備え、クロック信号の第１のタイミングで、入力オフセットアドレスレジスタに値をセットし、クロック信号の第２のタイミングで、入力オフセットアドレスレジスタの値をデータ保持レジスタに伝搬させ、クロック信号の第３のタイミングで、各々のアドレス加算結果と入力オフセットアドレスレジスタの値を加算し、クロック信号の第４のタイミングで、アドレスマスク機構により互いに異なる空間に分離してローカルメモリ参照のためのアドレスラッチに格納し、クロック信号の第５のタイミングで、ローカルメモリから読み出したインデックスとデータの組をメモリ出力レジスタに格納し、クロック信号の次のタイミングで、メモリ出力レジスタから読み出し、インデックス値として利用する。

[0014] 本発明の第１の観点の演算方法は、３入力パイプライン浮動小数点積和演算器とローカルメモリから構成される演算ユニットを用いる演算方法であって、インデックスレジスタとデータレジスタを組にした第１の入力レジスタ及び第２の入力レジスタ、データレジスタである第３の入力レジスタ及び出力レジスタを備える積和演算器を用い、下記１－１）～１－３）の各ステップを備える。

１－１）積和演算器の第１の入力レジスタ及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納されるステップ。

１－２）インデックス値Ａとインデックス値Ｂが比較されるステップ。

１－３）インデックス値Ａとインデックス値Ｂが一致する場合には、データＡとデータＢの乗算値と第３の入力レジスタのデータ値が加算されて積和演算器の出力レジスタに格納されると共に、第３の入力レジスタに戻すステップ。

[0015] 本発明の第１の観点の演算方法では、ＣＧＲＡ型アクセラレータに、データパスを追加して、少ないローカルメモリを用いた疎行列－疎行列の行列積の演算を行う。

行列の要素番号であるインデックス値を上位３２ｂｉｔ、行列の要素値であるデータを下位３２ｂｉｔ格納する６４ｂｉｔデータをローカルメモリに格納し、疎行列－疎行列の行列積の演算を行う。２つの入力が、疎行列であることに着目して、零要素を取り除き、データを圧縮して計算を行い、疎行列－疎行列の積を、疎行列－密行列を同じスループットで高効率に計算する。ＣＧＲＡはパイプライン型垂直方向であり、垂直方向（縦方向）の圧縮はできないことから、水平方向（横方向）の零要素を取り除きデータを圧縮し、データパスを追加することにより、少ないローカルメモリを用いた疎行列計算を途切れることなく行う。

[0016] 本発明の第１の観点の演算方法は、疎行列Ａと疎行列Ｂの行列積の演算方法であって、下記１－４），１－５）の各ステップを更に備えることが好ましい。

１－４）インデックス値Ａは疎行列の列番号で、インデックス値Ｂは疎行列の転置行列の列番号で、各疎行列におけるゼロ値の要素が圧縮されるステップ。

１－５）各疎行列における非ゼロ値の要素がインデックス値とデータを組と

してローカルメモリに記憶されるステップ。

[0017] 疎行列におけるゼロ値の要素の圧縮は、ゼロ値のものを取り除き、非ゼロ値のものだけを残して、列番号を前に詰め、インデックス値には元の列番号が格納され、データには元の列番号の値が格納される。

[0018] 本発明の第１の観点の演算方法は、下記１－６）～１－９）の各ステップを更に備えることが好ましい。

１－６）各々の入力ベースアドレスレジスタおよび入力オフセットアドレスレジスタにアドレス情報をロードするステップ。

１－７）ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂを比較するステップ。

１－８）第１のアドレス計算機構が、Ａ≦Ｂの場合に入力ベースアドレスレジスタに次要素参照に必要な値を加算するステップ、或いは、第２のアドレス計算機構が、Ａ≧Ｂの場合に入力ベースアドレスレジスタに次要素参照に必要な値を加算するステップ。

１－９）アドレス加算結果を入力ベースアドレスレジスタに戻すステップ。

[0019] 次に、本発明の第２の観点の演算方法は、３入力パイプライン浮動小数点積和演算器とローカルメモリから構成される演算ユニットを用いるマージソートの演算方法であって、インデックスレジスタとデータレジスタを組にした第１の入力レジスタ及び第２の入力レジスタ、データレジスタである第３の入力レジスタ及び出力レジスタを備える積和演算器を用い、下記２－１）～２－５）の各ステップを備える。

２－１）積和演算器の第１の入力レジスタ及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納されるステップ。

２－２）インデックス値Ａとインデックス値Ｂが比較されるステップ。

２－３）インデックス値Ａとインデックス値Ｂのデータの読み出し結果の大小関係に従い、後続の演算ユニットにローカルメモリの２つの読み出しアド

レスと２つの読み出しデータを送り、アドレスとデータの各々の大小関係に従い、何れかのデータをストアするステップ。

２－４）マージソート全体のＬｏｇＮ段のうち、１段分のソート結果をローカルメモリにストアするステップ。

２－５）ローカルメモリのストア先アドレスを単調に増加するステップ。

[0020] 本発明の第２の観点の演算方法は、ＣＧＲＡ型アクセラレータに、データパスを追加して、少ないローカルメモリを用いて、データ全体を小さい又は大きい順にデータをソートしマージソートを行う。

ローカルメモリに格納された６４ｂｉｔデータの内、２つの入力の各インデックス値である各上位３２ｂｉｔのデータの読み出し結果の大小関係に従い、マージソートする。

発明の効果

[0021] 本発明のＣＧＲＡによる演算ユニットによれば、少ないローカルメモリを用いて疎行列－疎行列の積、マージソートが可能となるといった効果がある。

図面の簡単な説明

[0022] [図1]疎行列－疎行列の積をＣＧＲＡに実装する方法の説明図

[図2]疎行列の圧縮方法の説明図

[図3]従来の演算ユニットにおけるデータの足し合わせの説明図

[図4]本発明の演算ユニットにおけるデータ循環の説明図

[図5]本発明の演算ユニットの構成模式図

[図6]演算ユニットの構成図

[図7]アドレス計算機構の構成図

[図8]実施例１の演算ユニットの動作説明図（１）

[図9]実施例１の演算ユニットの動作説明図（２）

[図10]実施例１の演算ユニットの動作説明図（３）

[図11]実施例１の演算ユニットの動作説明図（４）

[図12]実施例１の演算ユニットの動作説明図（５）

[図13]実施例２の演算ユニットの動作説明図（１）

[図14]実施例２の演算ユニットの動作説明図（２）

[図15]実施例１と実施例２のアドレス計算部分の共通性についての説明図

[図16]疎行列積の演算フロー図

[図17]マージソートの演算フロー図

発明を実施するための形態

[0023] 以下、本発明の実施形態の一例を、図面を参照しながら詳細に説明していく。なお、本発明の範囲は、以下の実施例や図示例に限定されるものではなく、幾多の変更及び変形が可能である。

[0024] まず、疎行列－疎行列の積をＣＧＲＡに実装する方法について、図１を参照しながら説明する。ここでは、図１に示すように、４行８列の行列Ａと８行４列の行列Ｂを掛け合わせて、４行４列の行列Ｃを計算する場合で説明する。行列Ａと行列Ｂの行列積を計算する場合に、定義通りに計算すると、枠で囲むように、行方向と列方向を掛け合わせることになるが（図１（１）を参照）、これではメモリアクセスが順番にならないことから、行列Ｂの転置行列B^Tを用いてメモリアクセスが順番になるようにする（図１（２）を参照）。

[0025] 行列Ａ、行列Ｂが共に疎行列、すなわち、行列の要素の多くがゼロ値で構成される行列である場合には、非ゼロ値の要素のみを選択して計算することにより計算量を削減できる。例えば、図１（３）に示すように、行列Ａの１行目の計算では、非ゼロ値のＡ００，Ａ０１だけを計算すればよく、その他のゼロ値の要素の計算はスキップできる。行列Ａの１行目のＡ００，Ａ０１と掛け合わせる転置行列B^Tの１～４行は、それぞれ、１行目はＢ００，Ｂ１０、２行目はＢ０１，Ｂ１１，Ｂ２１，Ｂ３１、３行目はＢ４２，Ｂ５２，Ｂ６２，Ｂ７２、４行目はＢ６３，Ｂ７３となる。

[0026] そして、図１（４）に示す行列Ａと転置行列B^Tを圧縮する。

図２に示すように、圧縮は、行列Ａと転置行列B^Tにおける非ゼロ値の要素をインデックス値（行列の要素番号）とデータ（要素値）の組に置き換えて

、ゼロ値の要素を全て取り除く。元の行列のサイズを維持するため、残りのインデックス値とデータの組は、インデックス値として負の値（例えば－１）をセットする。

[0027] 図３に示す従来の演算では、例えば、行列Ａの要素Ａ００が、行列Ｂの要素Ｂ００，Ｂ０１，Ｂ０２，Ｂ０３と掛け合わされたデータが、縦方向に流れ、後段の行列Ａの要素Ａ０１が、行列Ｂの要素Ｂ１０，Ｂ１１，Ｂ１２，Ｂ１３と掛け合わされたデータと足し合わされ、同様に、行列Ａの要素Ａ０２，Ａ０３が、それぞれ、行列Ｂの要素Ｂ２０～Ｂ２３、Ｂ３０～Ｂ３３と掛け合わされたデータと足し合わされ、最後に、行列Ｃの要素Ｃ００，Ｃ０１，Ｃ０２，Ｃ０４となる。このように、データの足し合わせは、縦方向と同じように演算ユニットを並べるが、図１（４）に示す行列Ａと転置行列B^Tを圧縮したものでは、縦方向と同じように演算ユニットをつなぎ合わせて実行することは不可能である。

[0028] そのため、図４に示すとおり、行列Ａと転置行列B^Tを圧縮したものは、１つの演算ユニットの内部でインデックス値を参照して疎行列のデータを循環させて、データの足し合わせを行う必要がある。

[0029] 以下では、疎行列Ａと疎行列Ｂの転置行列B^Tとを圧縮したもの（圧縮済み疎行列）を外部メモリから取り出して、１つの演算ユニットの内部でインデックス値を参照して疎行列のデータを循環させて、データの足し合わせを行う演算ユニットについて説明を行う。

[0030] 図５は、本発明の演算ユニットの構成模式図を示す。本発明の演算ユニットは、第１のアドレス計算機構１と第２のアドレス計算機構と演算ユニット３から構成され、それぞれが内部に持つローカルメモリ（１５，２５，３５）は外部メモリ５と接続されている。

第１のアドレス計算機構１の演算器１０では、入力ベースアドレスＡと、疎行列Ａと転置行列B^Tのインデックス値が入力され、各々のインデックス値を比較し、Ｂ≧Ａである場合には、インデックス値Ａのアドレスが増加され、疎行列Ａの次要素の値が参照される。次要素のインデックス値は、インデ

ックス値とデータが組となり、例えば、６４ｂｉｔ（＝３２ｂｉｔ＋３２ｂｉｔ）離れて記憶されているとすると、＋８ｂｙｔｅアドレスを増加する。一方、各々のインデックス値を比較し、Ｂ≧Ａでない場合には、インデックス値Ａのアドレスはそのまま維持し、アドレスは増加しない。

また、演算器（加算器）１１では、演算器１０が出力するアドレスに、入力オフセットアドレスＡの値が加算される。ローカルメモリ１５には、疎行列Ａのインデックス値とデータの組が記憶されている。メモリ出力レジスタ１６は、ローカルメモリ１５から読み出したインデックス値とデータの組を保持する。

[0031] 第２のアドレス計算機構２の演算器２０では、入力ベースアドレスＢと、疎行列Ａと転置行列B^Tのインデックス値が入力され、各々のインデックス値を比較し、Ｂ≦Ａである場合には、インデックス値Ｂのアドレスが増加され、転置行列B^Tの次要素の値が参照される。次要素のインデックス値は、インデックス値とデータが組となり、例えば、６４ｂｉｔ（＝３２ｂｉｔ＋３２ｂｉｔ）離れて記憶されているとすると、＋８ｂｙｔｅアドレスを増加する。一方、各々のインデックス値を比較し、Ｂ≦Ａでない場合には、インデックス値Ｂのアドレスはそのまま維持し、アドレスは増加しない。

また、演算器（加算器）２１では、演算器２０が出力するアドレスに、入力オフセットアドレスＢの値が加算される。ローカルメモリ２５には、転置行列B^Tのインデックス値とデータの組が記憶されている。メモリ出力レジスタ２６は、ローカルメモリ２５から読み出したインデックス値とデータの組を保持する。

[0032] 演算ユニット３は、３入力パイプライン浮動小数点積和演算器３０とローカルメモリ３５から構成される。演算器３０は、第１及び第２のアドレス計算機構から出力された第１の入力レジスタ３１と第２の入力レジスタ３２と前回のプロセスの出力結果（第３の入力レジスタ３３）とから３入力を行う。ローカルメモリ３５は、外部メモリ５と接続される。

[0033] 図５に示す本発明の演算ユニットでは、下記（１）～（５）の４段パイプ

ライン処理の５つのデータフローが連携して途切れなく演算を継続する。４段パイプライン処理は、４段階で一つの命令の処理が終わるようにし、それぞれの処理が独立に、かつ、同時に処理され、並行して動作できる処理である。

[0034] （１）浮動小数点積和演算器とローカルメモリのデータフロー（４段パイプライン処理）

クロック信号のタイミングｎ＋１で、積和演算器の第１段の第１の入力レジスタ及び第２の入力レジスタに、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納される。

クロック信号のタイミングｎ＋２で、インデックス値Ａとインデックス値Ｂが比較され、値が異なる場合には定数Ｃが積和演算器の第１段の出力データレジスタに格納され、値が一致する場合にはデータＡとデータＢの乗算が実行されて積和演算器の第１段の出力データレジスタに格納される。

クロック信号のタイミングｎ＋３で、積和演算器の第２段の出力データレジスタに加算結果が伝搬される。クロック信号のタイミングｎ＋４で、積和演算器の第３段の出力データレジスタに正規化済の積和演算結果が伝搬される。先頭へ戻るデータ累算リングが構成される。

[0035] （２）第１のアドレス計算機構のアドレス計算のデータフロー（４段パイプライン処理）

クロック信号の第１のタイミングで、入力ベースアドレスレジスタＡおよび入力オフセットアドレスレジスタＡにアドレス情報をロードする。

クロック信号の第２のタイミングで、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂを比較し、入力ベースアドレスレジスタＡに次要素参照に必要な値（例えば、＋８ｂｙｔｅ）を加算するか、或いは、そのままの値を維持する（＋０ｂｙｔｅ）。

クロック信号の第３および第４のタイミングは通過し、先頭へ戻るアドレ

ス累算リングが構成される。

[0036] （３）第２のアドレス計算機構のアドレス計算のデータフロー（４段パイプライン処理）

クロック信号の第１のタイミングで、入力ベースアドレスレジスタＢおよび入力オフセットアドレスレジスタＢにアドレス情報をロードする。

クロック信号の第２のタイミングで、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂを比較し、入力ベースアドレスレジスタＢに次要素参照に必要な値（例えば、＋８ｂｙｔｅ）を加算するか、或いは、そのままの値を維持する（＋０ｂｙｔｅ）。

クロック信号の第３および第４のタイミングは通過し、先頭へ戻るアドレス累算リングが構成される。

[0037] （４）第１のアドレス計算機構のデータ参照フロー（４段パイプライン処理）

クロック信号の第２のタイミングで、ローカルメモリから読み出したインデックス値Ａ，Ｂを比較し、入力ベースアドレスレジスタＡに＋０ｂｙｔｅ又は＋８ｂｙｔｅを加算出力し、

クロック信号の第３のタイミングで、入力オフセットアドレスレジスタＡを加算し、

クロック信号の第４のタイミングで、アドレスマスク機構により互いに異なる空間に分離してローカルメモリ参照のためのアドレスラッチに格納し、

クロック信号の第５のタイミングで、ローカルメモリから読み出したインデックスとデータの組をメモリ出力レジスタに格納し、先頭に戻る。

[0038] （５）第２のアドレス計算機構のデータ参照フロー（４段パイプライン処理）

クロック信号の第２のタイミングで、ローカルメモリから読み出したインデックス値Ａ，Ｂを比較し、入力ベースアドレスレジスタＢに＋０ｂｙｔｅ又は＋８ｂｙｔｅを加算出力し、

クロック信号の第３のタイミングで、入力オフセットアドレスレジスタＢ

を加算し、

クロック信号の第４のタイミングで、アドレスマスク機構により互いに異なる空間に分離してローカルメモリ参照のためのアドレスラッチに格納し、

クロック信号の第５のタイミングで、ローカルメモリから読み出したインデックスとデータの組をメモリ出力レジスタに格納し、先頭に戻る。

[0039]（演算ユニットの動作）

図６を参照して、本発明の演算ユニットの動作について説明する。

本発明の演算ユニットは、３入力パイプライン浮動小数点積和演算器３０とローカルメモリ３５から構成され、積和演算器３０は、インデックスレジスタ３１２とデータレジスタ３１１を組にした第１の入力レジスタ３１、インデックスレジスタ３２２とデータレジスタ３２１を組にした第２の入力レジスタ３２、データレジスタである第３の入力レジスタ３３、及び、出力レジスタ３４を備える。６４ｂｉｔデータで構成される第１の入力レジスタ３１には、上位３２ｂｉｔのインデックスレジスタ３１２に疎行列Ａの要素番号が、下位３２ｂｉｔのデータレジスタ３１１に要素値が格納される。また、６４ｂｉｔデータで構成される第２の入力レジスタ３２には、上位３２ｂｉｔのインデックスレジスタ３２２に疎行列Ｂの転置行列B^Tの要素番号が、下位３２ｂｉｔのデータレジスタ３２１に要素値が格納される。

[0040] 具体的には、積和演算器３０の第１の入力レジスタ３１及び第２の入力レジスタ３２には、図示しないローカルメモリ（行列積Ａ，行列積Ｂ）から読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタ（３１２，３２２）に格納され、データＡとデータＢが各データレジスタ（３１１，３２１）に格納される。

そして、インデックス値Ａとインデックス値Ｂが比較され、値が一致する場合には、データＡとデータＢの乗算値と第３の入力レジスタ３３のデータ値（前回のプロセスのデータ結果）が加算されて積和演算器３０の出力レジスタ３４に格納され、かつ、第３の入力レジスタ３３に戻る。インデックス値Ａとインデックス値Ｂの値が一致しない場合には、データＡとデータＢの

乗算をしないでスキップし、第３の入力レジスタ３３のデータ値（前回のプロセスのデータ結果）をそのまま出力レジスタ３４に出力する。

[0041] このように、１つの演算ユニットの内部で２つのインデックス値Ａ，Ｂを参照し、疎行列のデータを出力から入力に戻し循環させて、疎行列積のデータの足し合わせを行っている。

そして、出力レジスタ３４の内容をローカルメモリ３５の行列積Ｃの各要素のアドレス領域に格納する。ローカルメモリ３５と外部メモリ５は接続され、行列積Ｃの各要素は外部メモリ５に保存される。

[0042] （アドレス計算機構の動作）

次に、図７を参照して、本発明のアドレス計算機構の動作について説明する。

本発明のアドレス計算機構は、入力ベースアドレスレジスタ（１８，２８）と入力オフセットアドレスレジスタ（１９，２９）を含む第１のアドレス計算機構および第２のアドレス計算機構を備える。

[0043] 第１のアドレス計算機構は、クロック信号の第１のタイミングで、入力ベースアドレスレジスタＡ１８および入力オフセットアドレスレジスタＡ１９にアドレス情報をロードする。クロック信号の第２のタイミングで、ローカルメモリ１５から読み出したインデックス値Ａとインデックス値Ｂを比較し、Ａ≦Ｂの場合に入力ベースアドレスレジスタＡ１８に次要素参照に必要な値（＋８ｂｙｔｅ）を加算する。クロック信号の第３および第４のタイミングで、アドレス加算結果をデータ保持レジスタＡ１４に伝搬させ、クロック信号の次のタイミングで、データ保持レジスタＡ１４の内容を入力ベースアドレスレジスタＡ１８に戻す。

[0044] 第２のアドレス計算機構は、クロック信号の第１のタイミングで、入力ベースアドレスレジスタＢ２８および入力オフセットアドレスレジスタＢ２９にアドレス情報をロードする。クロック信号の第２のタイミングで、ローカルメモリ２５から読み出したインデックス値Ａとインデックス値Ｂを比較し、Ａ≧Ｂの場合に入力ベースアドレスレジスタＢ２８に次要素参照に必要な

値（＋８ｂｙｔｅ）を加算する。クロック信号の第３および第４のタイミングで、アドレス加算結果をデータ保持レジスタＢ２４に伝搬させ、クロック信号の次のタイミングで、データ保持レジスタＢ２４の内容を入力ベースアドレスレジスタＢ２８に戻す。

[0045] 第１及び第２のアドレス計算機構は、それぞれ、ローカルメモリを複数空間に分割するアドレスマスク機構（１０１，２０１）を備える。

クロック信号の第１のタイミングで、入力オフセットアドレスレジスタ（１９，２９）に値をセットする。クロック信号の第２のタイミングで、入力オフセットアドレスレジスタ（１９，２９）の値をデータ保持レジスタ（図示せず）に伝搬させる。クロック信号の第３のタイミングで、各々のアドレス加算結果と入力オフセットアドレスレジスタの値を加算する。

[0046] クロック信号の第４のタイミングで、アドレスマスク機構により互いに異なる空間に分離してローカルメモリ参照のためのアドレスラッチに格納する。クロック信号の第５のタイミングで、ローカルメモリから読み出したインデックス値とデータの組をメモリ出力レジスタ（１６，２６）に格納する。クロック信号の次のタイミングで、メモリ出力レジスタ（１６，２６）から読み出し、インデックス値として利用する。

実施例 １

[0047] ４クロック単位で動作するＣＧＲＡによる疎行列計算に関する演算ユニットの一実施形態について、具体的に説明する。以下の説明で参照する図８～１２において、図中に数字が付されているが、下記説明におけるレジスタ番号と対応する（第ｎのレジスタは、図中のマルで囲んだ数字ｎに対応する）。

[0048] 演算ユニットと複数のアドレス生成器（アドレス計算機構）とローカルメモリの組から構成される基本ユニットを複数連結する。各アドレス生成器の第１ステージは、ベースアドレスを保持する第１のレジスタ（図７における入力ベースアドレスレジスタ１８，２８）の値と、自身が最終ステージにおいて生成するアドレス計算結果をアドレス生成器毎に保持する第９のレジス

タ（図７におけるデータ保持レジスタ１４，２４）の値と、ローカルメモリの２カ所から読み出された２つのデータを各々保持する第３および第４のレジスタ（図７におけるメモリ出力レジスタ１６，２６）の値とを入力する。

[0049] 初回のループイタレーションでは、各アドレス生成器の第１ステージは第１のレジスタ（入力ベースレジスタ１８，２８）の値を加算の第１オペランド、定数０を加算の第２オペランドとして加算結果を第２のレジスタに格納する。

２回目以降のループイタレーションでは、各アドレス生成器の第１ステージはアドレス生成器毎の第９のレジスタ（データ保持レジスタ１４，２４）の値を第１オペランド、第３および第４のレジスタ（メモリ出力レジスタ１６，２６）の各上位ｂｉｔの比較結果によりセットされる定数を第２オペランドして加算結果を各アドレス生成器の第２のレジスタに格納する。

[0050] 各アドレス生成器の第２ステージは、第２のレジスタの値を加算の第１オペランド、オフセットアドレスを保持する第５のレジスタ（入力オフセットアドレスレジスタ１９，２９）の値を加算の第２オペランドとして加算を行い、加算結果を第６のレジスタに格納し、同時に、第２のレジスタの内容をそのまま第７のレジスタに格納する。

[0051] 各アドレス生成器の第３ステージは、第６のレジスタの値をローカルメモリ参照用マスク操作の第１オペランド、マスク値を第２オペランドとしてマスク操作を行い、マスク結果を各アドレス生成器の第８のレジスタに格納し、同時に、第７のレジスタの内容をそのまま各アドレス生成器の第９のレジスタ（データ保持レジスタ１４，２４）に格納する。

[0052] 各アドレス生成器が生成し第８のレジスタに格納するアドレス情報を用いて、ローカルメモリの複数個所を参照し、ロード結果を第３および第４のレジスタ（メモリ出力レジスタ１６，２６）に各々格納すると同時に、各々演算ユニットの境界に配置されるレジスタファイル中の第１０および第１１のレジスタ（図５，図６における第１及び第２の入力レジスタ３１，３２）に格納する。

[0053] 第１０および第１１のレジスタ（第１及び第２の入力レジスタ３１，３２）の各下位３２ｂｉｔのデータＡ，Ｂを演算オペランドとして演算ユニット３０に入力する。

実施例 2

[0054] ４クロック単位で動作するＣＧＲＡによるマージソートに関する演算ユニットの一実施形態について、具体的に説明する。以下の説明で参照する図１３～１５において、図中に数字が付されているが、下記説明におけるレジスタ番号と対応する（第ｎのレジスタは、図中のマルで囲んだ数字ｎに対応する）。

マージソートとは、データ全体を小さい又は大きい順にデータをソートしマージする演算である。図１３に示すとおり、実施例１の疎行列計算と同様に計算を行い、第１０および第１１のレジスタ（図６における第１及び第２の入力レジスタ３１，３２）の各上位３２ｂｉｔのデータの読み出し結果の大小関係に従い、１次元配列内の異なるアドレスＡ，Ｂの片側を更新してマージソートしていく。なお、実施例１で説明した疎行列計算に関する演算ユニットにおいて、上位３２ｂｉｔのインデックス値を比較するのと同様に、本実施例のマージソートに関する演算ユニットでも上位３２ｂｉｔをデータとして比較する。そして、マージソートの場合、第１０および第１１のレジスタの各下位３２ｂｉｔは、データの付随情報として、単にストアされるだけである。

[0055] 本実施例のマージソートに関する演算ユニットについても、実施例１の疎行列計算と同様に、演算ユニットと複数のアドレス生成器（アドレス計算機構）とローカルメモリの組から構成される基本ユニットを複数連結する。各アドレス生成器の第１ステージは、ベースアドレスを保持する第１のレジスタ（図７における入力ベースアドレスレジスタ１８，２８）の値と、自身が最終ステージにおいて生成するアドレス計算結果をアドレス生成器毎に保持する第９のレジスタ（図７におけるデータ保持レジスタ１４，２４）の値と、ローカルメモリの２カ所から読み出された２つのデータを各々保持する第

３および第４のレジスタ（図７におけるメモリ出力レジスタ１６，２６）の値とを入力する。

そして、初回のループイタレーションでは、各アドレス生成器の第１ステージは第１のレジスタ（入力ベースレジスタ１８，２８）の値を加算の第１オペランド、定数０を加算の第２オペランドとして加算結果を第２のレジスタに格納する。なお、２回目以降のループイタレーションでは、各アドレス生成器の第１ステージはアドレス生成器毎の第９のレジスタ（データ保持レジスタ１４，２４）の値を第１オペランド、第３および第４のレジスタ（メモリ出力レジスタ１６，１７）の各上位ｂｉｔの比較結果によりセットされる定数を第２オペランドして加算結果を各アドレス生成器の第２のレジスタに格納する。

[0056] 各アドレス生成器の第２ステージは、第２のレジスタの値を加算の第１オペランド、オフセットアドレスを保持する第５のレジスタ（入力オフセットアドレスレジスタ１９，２９）の値を加算の第２オペランドとして加算を行い、加算結果を第６のレジスタに格納し、同時に、第２のレジスタの内容をそのまま第７のレジスタに格納する。

次に、各アドレス生成器の第３ステージは、第６のレジスタの値をローカルメモリ参照用マスク操作の第１オペランド、マスク値を第２オペランドとしてマスク操作を行い、マスク結果を各アドレス生成器の第８のレジスタに格納し、同時に、第７のレジスタの内容をそのまま各アドレス生成器の第９のレジスタ（データ保持レジスタ１４，２４）に格納する。

そして、各アドレス生成器が生成し第８のレジスタに格納するアドレス情報を用いて、ローカルメモリの複数個所を参照し、ロード結果を第３および第４のレジスタ（メモリ出力レジスタ１６，２６）に各々格納すると同時に、各々演算ユニットの境界に配置されるレジスタファイル中の第１０および第１１のレジスタ（図５，図６における第１及び第２の入力レジスタ３１，３２）に格納する。

[0057] 第１０および第１１のレジスタ（第１及び第２の入力レジスタ３１，３２

）の各上位３２ｂｉｔのデータの読み出し結果の大小関係に従い、後続の演算ユニット３０にアドレスＡ，Ｂと、２つの読み出しデータを送り、アドレスとデータの各々の大小関係に従い、何れかのデータをストアすることにより、ソート全体のＬｏｇＮ段のうち、１段分のソート結果をローカルメモリにストアする。ストア先アドレスは単調増加していく。

同時に後続の演算ユニット３０が、前回の実行結果を前段のローカルメモリから読み出して、ＬｏｇＮ段の次段以降を担当することから、全体として、ローカルメモリをダブルバッファとするパイプライン実行が可能となるのである。

このように、本演算ユニットは、実施例１に示した疎行列積の演算の後、後続の演算ユニットにアドレスＡ，Ｂと、２つの読み出しデータを送り、アドレスとデータの各々の大小関係に従い、何れかのデータをストアさせることにより、マージソートの演算として使用できることがわかる。

[0058] 図１４は、４クロック単位で動作するＣＧＲＡの論理４列に写像された実際のデータフローを示している。

[0059] 図１５は、実施例１と実施例２の演算ユニットにおけるアドレス計算部分の共通性を示すべく、実際にＣ言語を用いたプログラムコードを示している。疎行列積の演算の場合（図１５におけるＡのコード部分）も、マージソートの演算の場合（図１５におけるＢのコード部分）も、プログラムコードの表現は同一であり、大小比較の部分の演算において、疎行列積の演算とマージソートの演算の何れの演算においてもアドレス計算部分の共通性があり、比較結果に基づいてアドレスを更新していることがわかる。

[0060] （その他の実施例）

（１）本発明の疎行列積の演算方法の一実施形態について、図１６の演算フローを参照して説明する。使用する演算ユニットは、３入力パイプライン浮動小数点積和演算器とローカルメモリから構成され、積和演算器は、インデックスレジスタとデータレジスタを組にした第１の入力レジスタ及び第２の入力レジスタ、データレジスタである第３の入力レジスタ及び出力レジスタ

を備える。

疎行列積の演算フローは、まず、積和演算器の第１及び第２の入力レジスタの各インデックスレジスタにインデックス値Ａとインデックス値Ｂを格納し、各データレジスタにデータＡとデータＢを格納する（ステップＳ０１）。インデックス値Ａは疎行列の列番号、インデックス値Ｂは疎行列の転置行列の列番号であり、各疎行列におけるゼロ値の要素を圧縮する（ステップＳ０２）。各疎行列における非ゼロ値の要素がインデックス値とデータを組としてローカルメモリに記憶する（ステップＳ０３）。次に、インデックス値Ａとインデックス値Ｂを比較する（ステップＳ０４）。

インデックス値Ａ，Ｂが一致する場合（Ａ＝Ｂ）、データＡとデータＢの乗算値と第３の入力レジスタのデータ値が加算されて積和演算器の出力レジスタに格納されると共に、第３の入力レジスタに戻す（ステップＳ０６）。

インデックス値Ａがインデックス値Ｂより小さいか等しい場合（Ａ≦Ｂ）、第１のアドレス計算機構が、入力ベースアドレスレジスタに次要素参照に必要な値を加算する（ステップＳ０７）。インデックス値Ａがインデックス値Ｂより大きいか等しい場合（Ａ≧Ｂ）、第２のアドレス計算機構が、入力ベースアドレスレジスタに次要素参照に必要な値を加算する（ステップＳ０８）。ステップＳ０７とステップＳ０８の場合に、その後、アドレス加算結果を入力ベースアドレスレジスタに戻す（ステップＳ０９）。これらのステップは、疎行列積が完了するまで繰り返され演算が完了する。

[0061] （２）本発明のマージソートの演算方法の一実施形態について、図１７の演算フローを参照して説明する。使用する演算ユニットは、３入力パイプライン浮動小数点積和演算器とローカルメモリから構成され、積和演算器は、インデックスレジスタとデータレジスタを組にした第１の入力レジスタ及び第２の入力レジスタ、データレジスタである第３の入力レジスタ及び出力レジスタを備える。

マージソートの演算フローは、まず、積和演算器の第１及び第２の入力レジスタの各インデックスレジスタにインデックス値Ａとインデックス値Ｂを

格納し、各データレジスタにデータＡとデータＢを格納する（ステップＳ１１）。次に、インデックス値Ａとインデックス値Ｂを比較する（ステップＳ１２）。そして、インデックス値Ａとインデックス値Ｂのデータの読み出し結果の大小関係に従い、後続の演算ユニットにローカルメモリの２つの読み出しアドレスと２つの読み出しデータを送り、アドレスとデータの各々の大小関係に従い、何れかのデータをローカルメモリにストアする（ステップＳ１３）。これにより、マージソート全体のＬｏｇＮ段のうち、１段分のソート結果をローカルメモリにストアされる（ステップＳ１４）。ローカルメモリのストア先アドレスを単調増加し（ステップＳ１５）、ステップＳ１１～Ｓ１５の処理が、マージソート全体のＬｏｇＮ段まで繰り返しされ、マージソート演算が完了する。

産業上の利用可能性

[0062] 本発明は、メモリ容量に制約がある超小型ＡＩアクセラレータに有用である。

符号の説明

[0063] １　第１のアドレス計算機構

２　第２のアドレス計算機構

３　演算ユニット

５　外部メモリ

３１，３２，３３　入力レジスタ

３４　出力レジスタ

１５，２５，３５　ローカルメモリ

請求の範囲

[請求項1] ３入力パイプライン浮動小数点積和演算器とローカルメモリから構成される演算ユニットであって、

積和演算器は、インデックスレジスタとデータレジスタを組にした第１の入力レジスタ及び第２の入力レジスタ、データレジスタである第３の入力レジスタ及び出力レジスタを備え、

積和演算器の第１の入力レジスタ及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納され、

インデックス値Ａとインデックス値Ｂが比較され、値が一致する場合には、データＡとデータＢの乗算値と第３の入力レジスタのデータ値が加算されて積和演算器の出力レジスタに格納されると共に、第３の入力レジスタに戻ることを特徴とする演算ユニット。

[請求項2] 疎行列Ａと疎行列Ｂの行列積の演算において、インデックス値Ａは疎行列の列番号で、インデックス値Ｂは疎行列の転置行列の列番号であり、

各疎行列におけるゼロ値の要素が圧縮され、非ゼロ値の要素がインデックス値とデータを組としてローカルメモリに記憶される請求項１に記載の演算ユニット。

[請求項3] クロック信号のタイミングｎ＋１で、積和演算器の第１段の第１の入力レジスタ及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納され、

クロック信号のタイミングｎ＋２で、インデックス値Ａとインデックス値Ｂが比較され、値が異なる場合には定数Ｃが積和演算器の第１段の出力データレジスタに格納され、値が一致する場合にはデータＡ

とデータＢの乗算が実行されて積和演算器の第１段の出力データレジスタに格納され、

クロック信号のタイミングｎ＋３で、積和演算器の第２段の出力データレジスタに加算結果が伝搬され、

クロック信号のタイミングｎ＋４で、積和演算器の第３段の出力データレジスタに正規化済の積和演算結果が伝搬され、

クロック信号のタイミングｎ＋５で、積和演算器の第３段の出力データレジスタから積和演算器の第１段の第３の入力レジスタであるデータレジスタに戻される、ことを特徴とする請求項１又は２に記載の演算ユニット。

[請求項4] 入力ベースアドレスレジスタと入力オフセットアドレスレジスタを含む第１のアドレス計算機構および第２のアドレス計算機構を備え、

第１および第２のアドレス計算機構は、各々の入力ベースアドレスレジスタおよび入力オフセットアドレスレジスタにアドレス情報をロードし、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂを比較し、第１のアドレス計算機構は、Ａ≦Ｂの場合に前記入力ベースアドレスレジスタに次要素参照に必要な値を加算し、第２のアドレス計算機構は、Ａ≧Ｂの場合に前記入力ベースアドレスレジスタに次要素参照に必要な値を加算し、アドレス加算結果を前記入力ベースアドレスレジスタに戻すことを特徴とする請求項１～３の何れかに記載の演算ユニット。

[請求項5] クロック信号の第１のタイミングで、各々の入力ベースアドレスレジスタおよび入力オフセットアドレスレジスタにアドレス情報をロードし、

クロック信号の第２のタイミングで、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂを比較し、第１のアドレス計算機構は、Ａ≦Ｂの場合に前記入力ベースアドレスレジスタに次要素参照に必要な値を加算し、第２のアドレス計算機構は、Ａ≧Ｂの場合

に前記入力ベースアドレスレジスタに次要素参照に必要な値を加算し、

クロック信号の第３および第４のタイミングで、アドレス加算結果をデータ保持レジスタに伝搬させ、

クロック信号の次のタイミングで、前記データ保持レジスタの内容を前記入力ベースアドレスレジスタに戻すことを特徴とする請求項４に記載の演算ユニット。

[請求項6] 第１及び第２のアドレス計算機構は、ローカルメモリを複数空間に分割するアドレスマスク機構を備え、

クロック信号の第１のタイミングで、前記入力オフセットアドレスレジスタに値をセットし、

クロック信号の第２のタイミングで、前記入力オフセットアドレスレジスタの値をデータ保持レジスタに伝搬させ、

クロック信号の第３のタイミングで、各々の前記アドレス加算結果と前記入力オフセットアドレスレジスタの値を加算し、

クロック信号の第４のタイミングで、前記アドレスマスク機構により互いに異なる空間に分離してローカルメモリ参照のためのアドレスラッチに格納し、

クロック信号の第５のタイミングで、ローカルメモリから読み出したインデックスとデータの組をメモリ出力レジスタに格納し、

クロック信号の次のタイミングで、前記メモリ出力レジスタから読み出し、前記インデックス値として利用することを特徴とする請求項５の演算ユニット。

[請求項7] ３入力パイプライン浮動小数点積和演算器とローカルメモリから構成される演算ユニットを用いる演算方法であって、

インデックスレジスタとデータレジスタを組にした第１の入力レジスタ及び第２の入力レジスタ、データレジスタである第３の入力レジスタ及び出力レジスタを備える積和演算器を用い、

積和演算器の第１の入力レジスタ及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納されるステップと、

インデックス値Ａとインデックス値Ｂが比較されるステップと、

インデックス値Ａとインデックス値Ｂが一致する場合には、データＡとデータＢの乗算値と第３の入力レジスタのデータ値が加算されて積和演算器の出力レジスタに格納されると共に、第３の入力レジスタに戻すステップ、を備えることを特徴とする演算方法。

[請求項8] 前記演算方法が、疎行列Ａと疎行列Ｂの行列積の演算方法であって、

インデックス値Ａは疎行列の列番号で、インデックス値Ｂは疎行列の転置行列の列番号で、各疎行列におけるゼロ値の要素が圧縮されるステップと、

各疎行列における非ゼロ値の要素がインデックス値とデータを組としてローカルメモリに記憶されるステップを更に備えることを特徴とする請求項７に記載の演算方法。

[請求項9] 各々の入力ベースアドレスレジスタおよび入力オフセットアドレスレジスタにアドレス情報をロードするステップと、

ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂを比較するステップと、

第１のアドレス計算機構が、Ａ≦Ｂの場合に前記入力ベースアドレスレジスタに次要素参照に必要な値を加算するステップ、或いは、第２のアドレス計算機構が、Ａ≧Ｂの場合に前記入力ベースアドレスレジスタに次要素参照に必要な値を加算するステップと、

アドレス加算結果を前記入力ベースアドレスレジスタに戻すステップを更に備えることを特徴とする請求項７又は８に記載の演算方法。

[請求項10] ３入力パイプライン浮動小数点積和演算器とローカルメモリから構

成される演算ユニットを用いるマージソートの演算方法であって、

インデックスレジスタとデータレジスタを組にした第１の入力レジスタ及び第２の入力レジスタ、データレジスタである第３の入力レジスタ及び出力レジスタを備える積和演算器を用い、

積和演算器の第１の入力レジスタ及び第２の入力レジスタには、ローカルメモリから読み出したインデックス値Ａとインデックス値Ｂが各インデックスレジスタに格納され、ローカルメモリから読み出したデータＡとデータＢが各データレジスタに格納されるステップと、

インデックス値Ａとインデックス値Ｂが比較されるステップと、

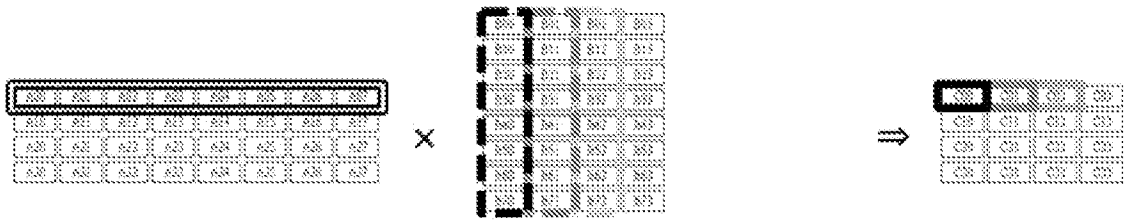
インデックス値Ａとインデックス値Ｂのデータの読み出し結果の大小関係に従い、後続の演算ユニットにローカルメモリの２つの読み出しアドレスと２つの読み出しデータを送り、アドレスとデータの各々の大小関係に従い、何れかのデータをストアするステップと、

マージソート全体のＬｏｇＮ段のうち、１段分のソート結果をローカルメモリにストアするステップと、

ローカルメモリのストア先アドレスを単調に増加するステップ、を備えることを特徴とする演算方法。

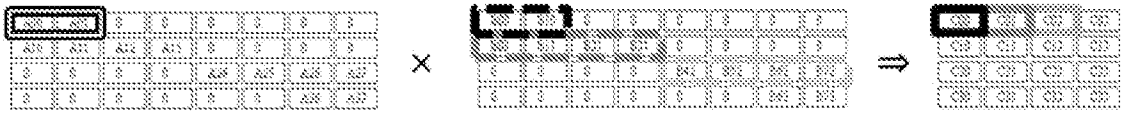
[図1]

(1) A＊B ⇒ C



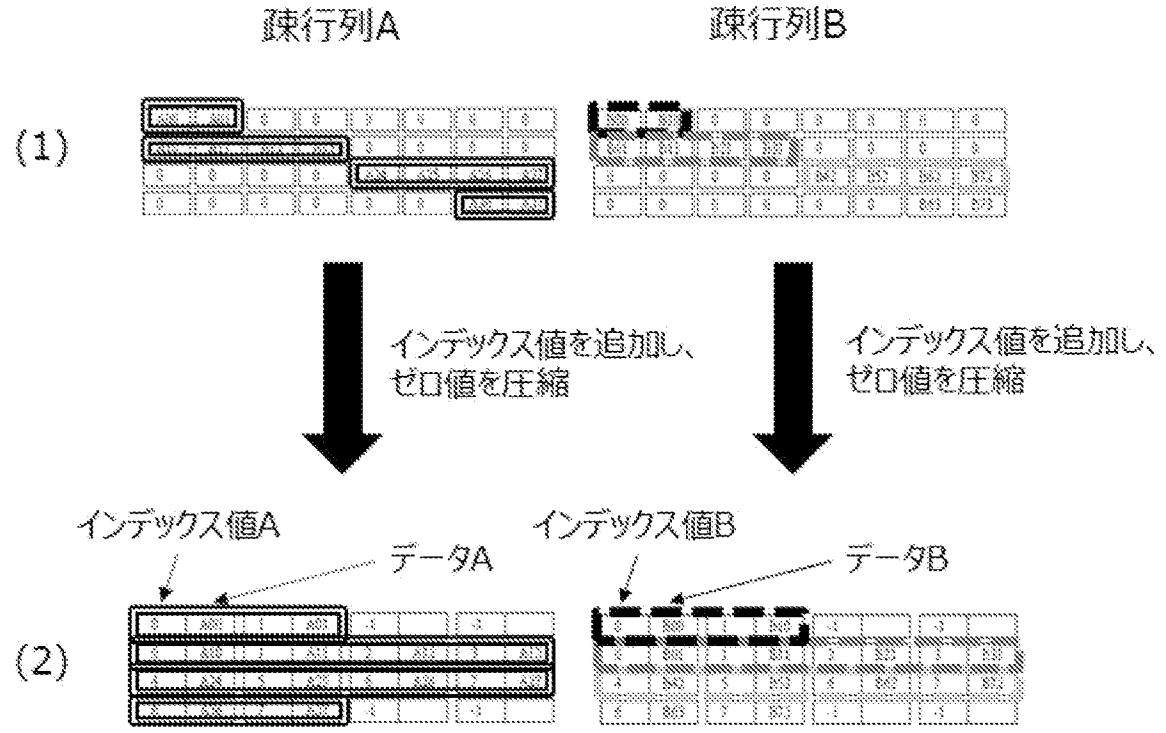
(2) A＊B^T ⇒ C

(3) A＊B^T(疎行列) ⇒ C

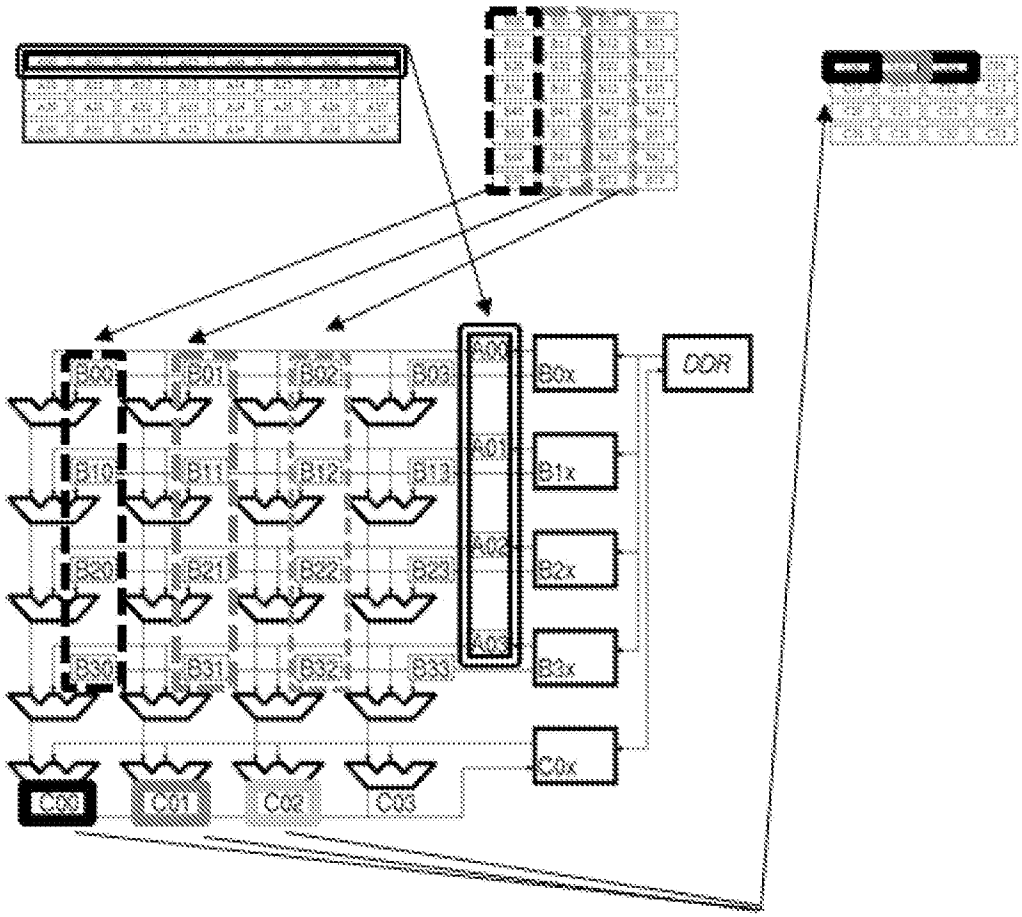


(4) A＊B^T(圧縮表現) ⇒ C

[図2]

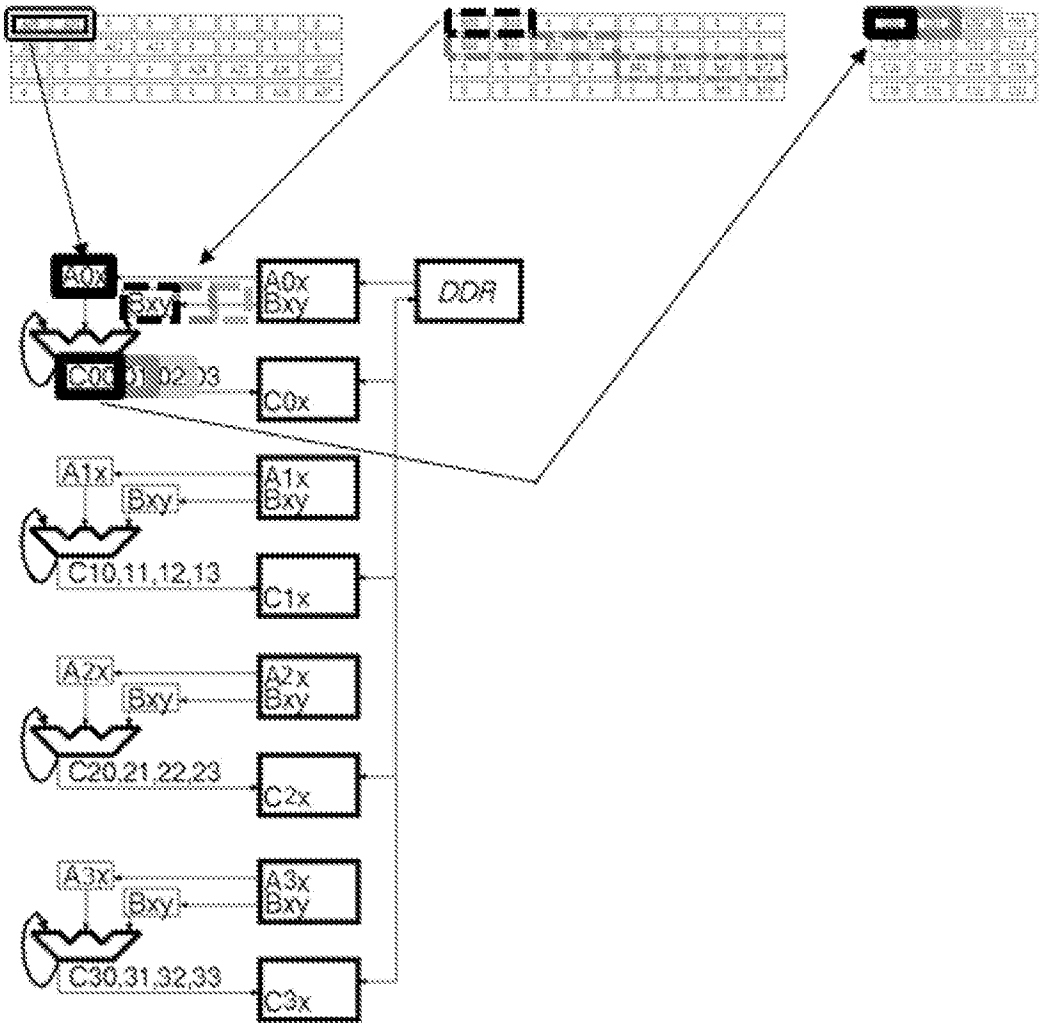


[図3]



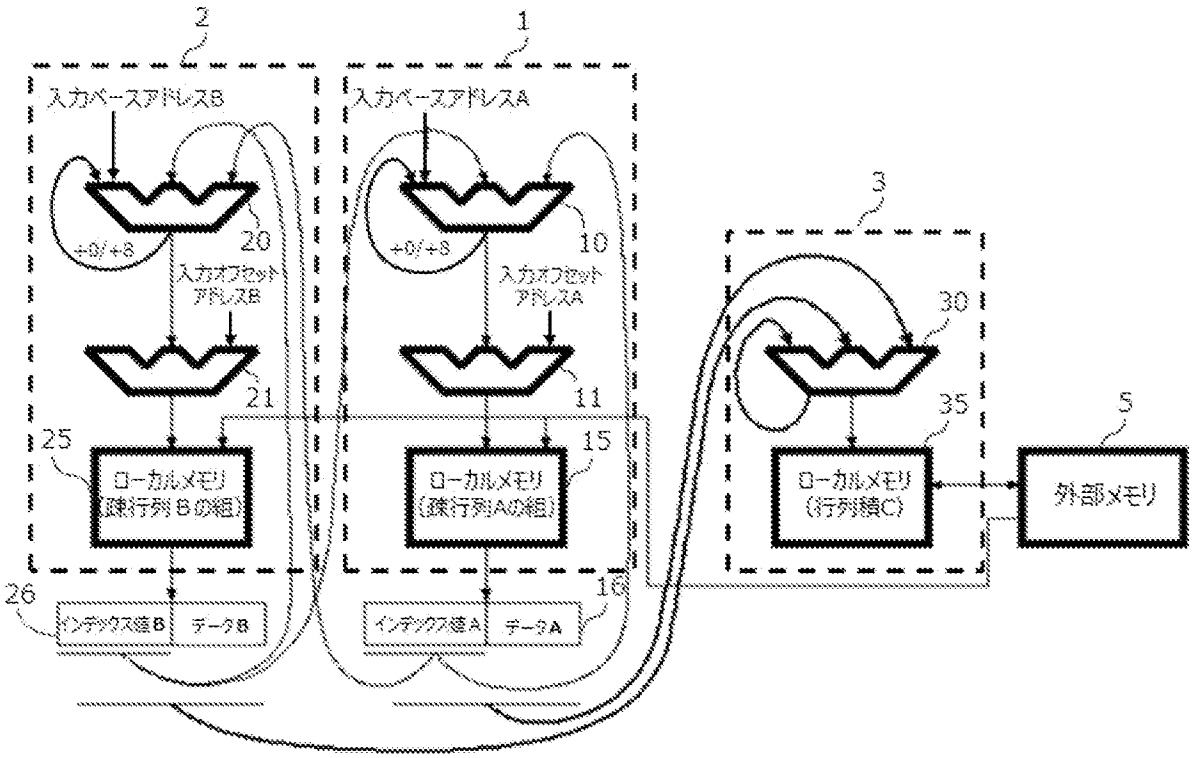
従来のCGRA演算の構成

[図4]

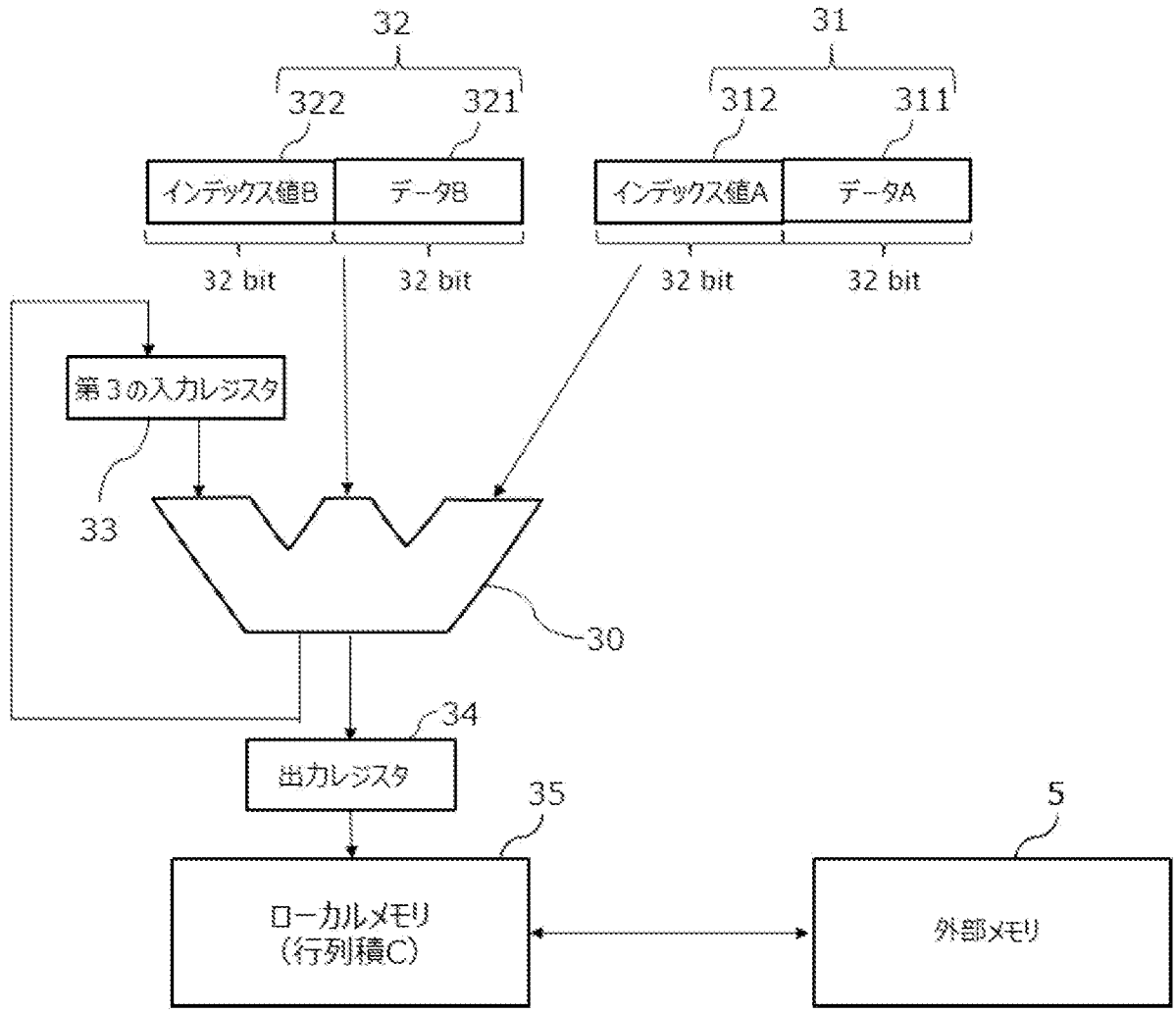


本発明のCGRA演算の構成

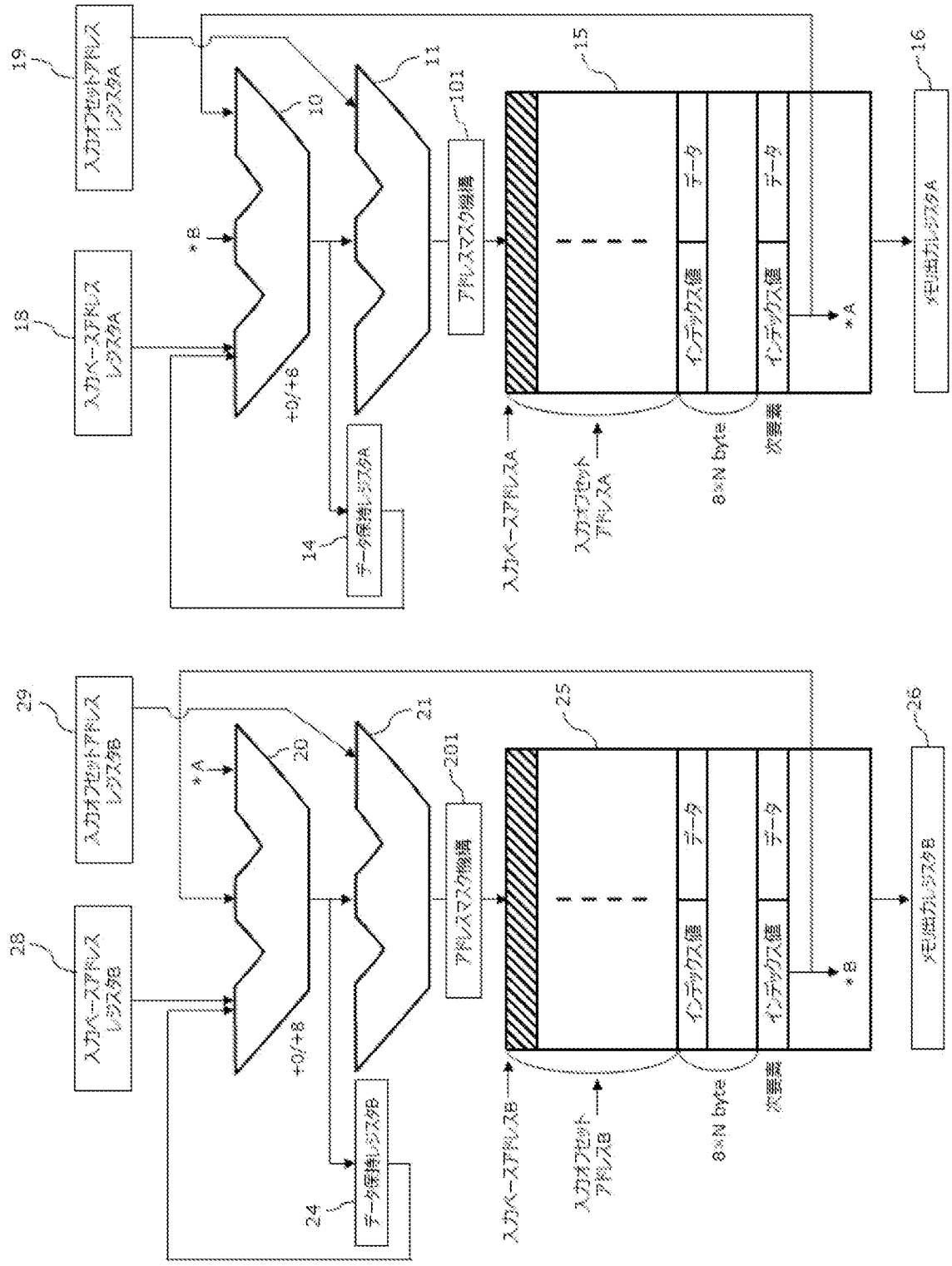
[図5]


2
1
入力ベースアドレスB
入力ベースアドレスA
3
+0/+8
20
+0/+8
10
入力オフセットアドレスB
入力オフセットアドレスA
30
21
11
35
5
25
ローカルメモリ（疎行列Bの組）
15
ローカルメモリ（疎行列Aの組）
ローカルメモリ（行列積C）
外部メモリ
26
インデックス値B
データB
16
インデックス値A
データA

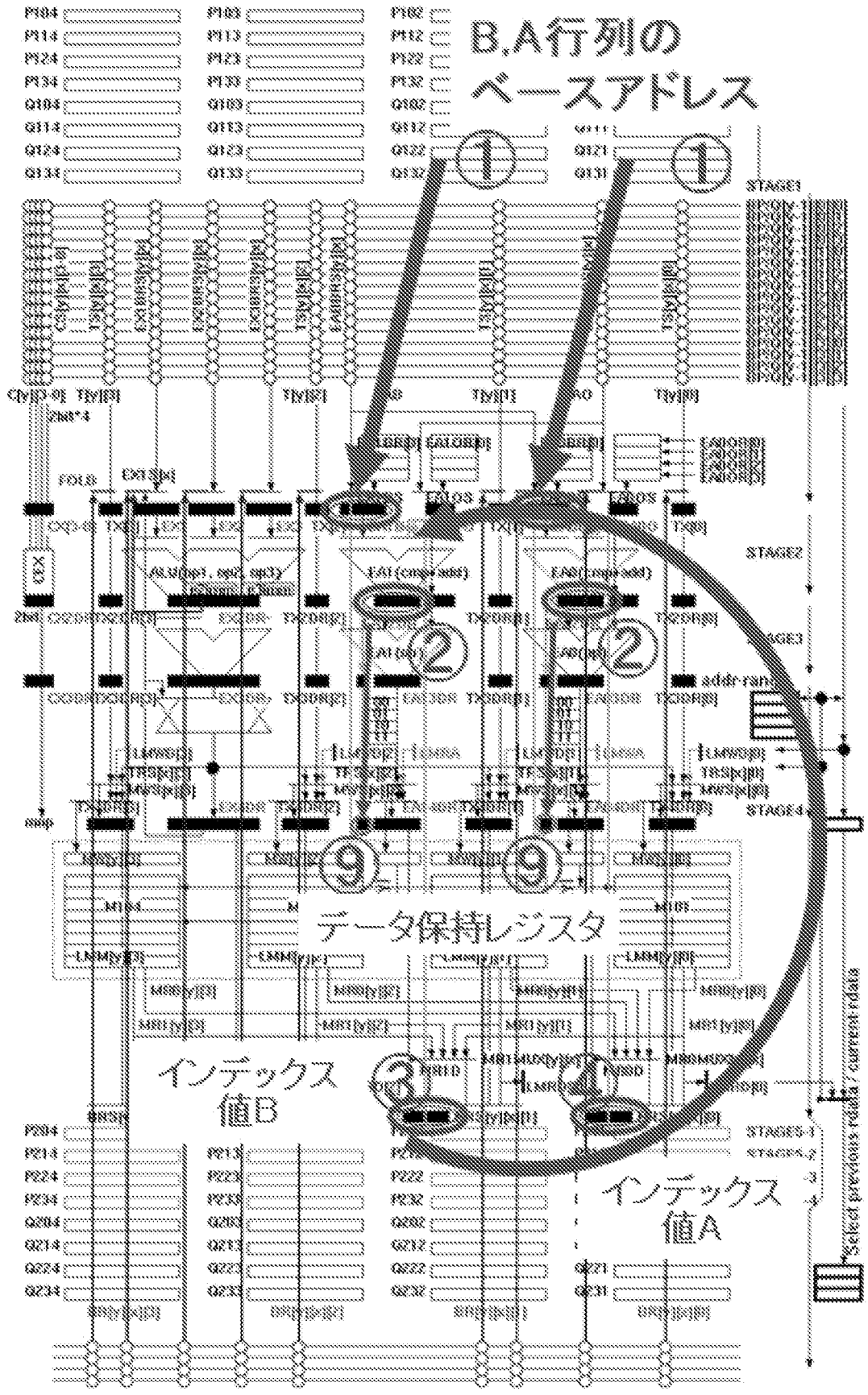
[図6]


32
31
322
321
312
311
インデックス値B
データB
インデックス値A
データA
32 bit
32 bit
32 bit
32 bit
第３の入力レジスタ
33
30
34
出力レジスタ
35
5
ローカルメモリ（行列積C）
外部メモリ

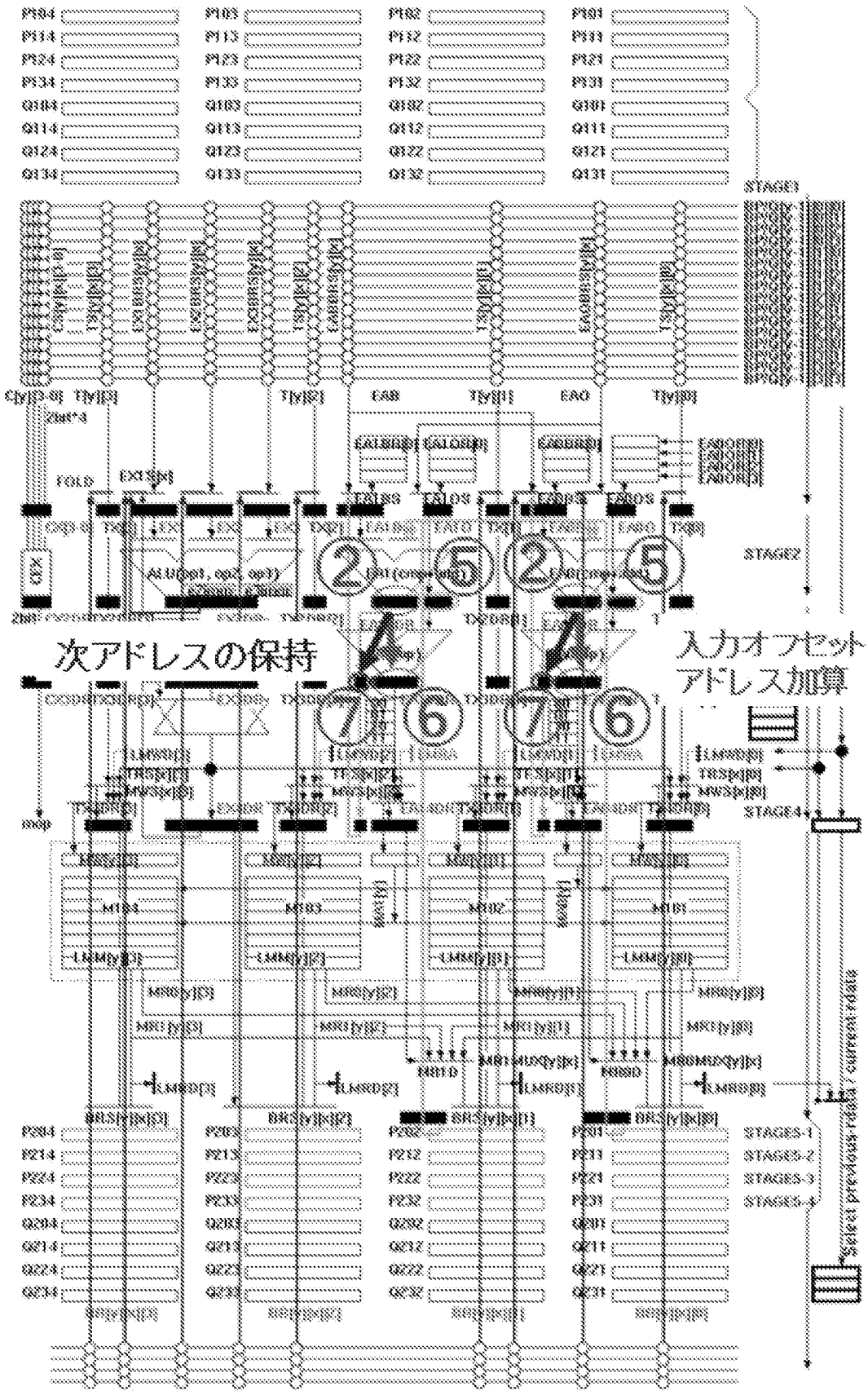
[図7]


19
ベースオフセットアドレスレジスタA
18
ベースアドレスレジスタA
10
11
101
アドレスマスク値
15
データ
インデックス値
16
実体化レジスタA
*B
+0/+8
14
データ保持レジスタA
ベースアドレスA
ベースオフセットアドレスA
8*N byte
次要素
*A
29
ベースオフセットアドレスレジスタB
28
ベースアドレスレジスタB
20
21
201
アドレスマスク値
25
データ
インデックス値
26
実体化レジスタB
*A
+0/+8
24
データ保持レジスタB
ベースアドレスB
ベースオフセットアドレスB
8*N byte
次要素
*B

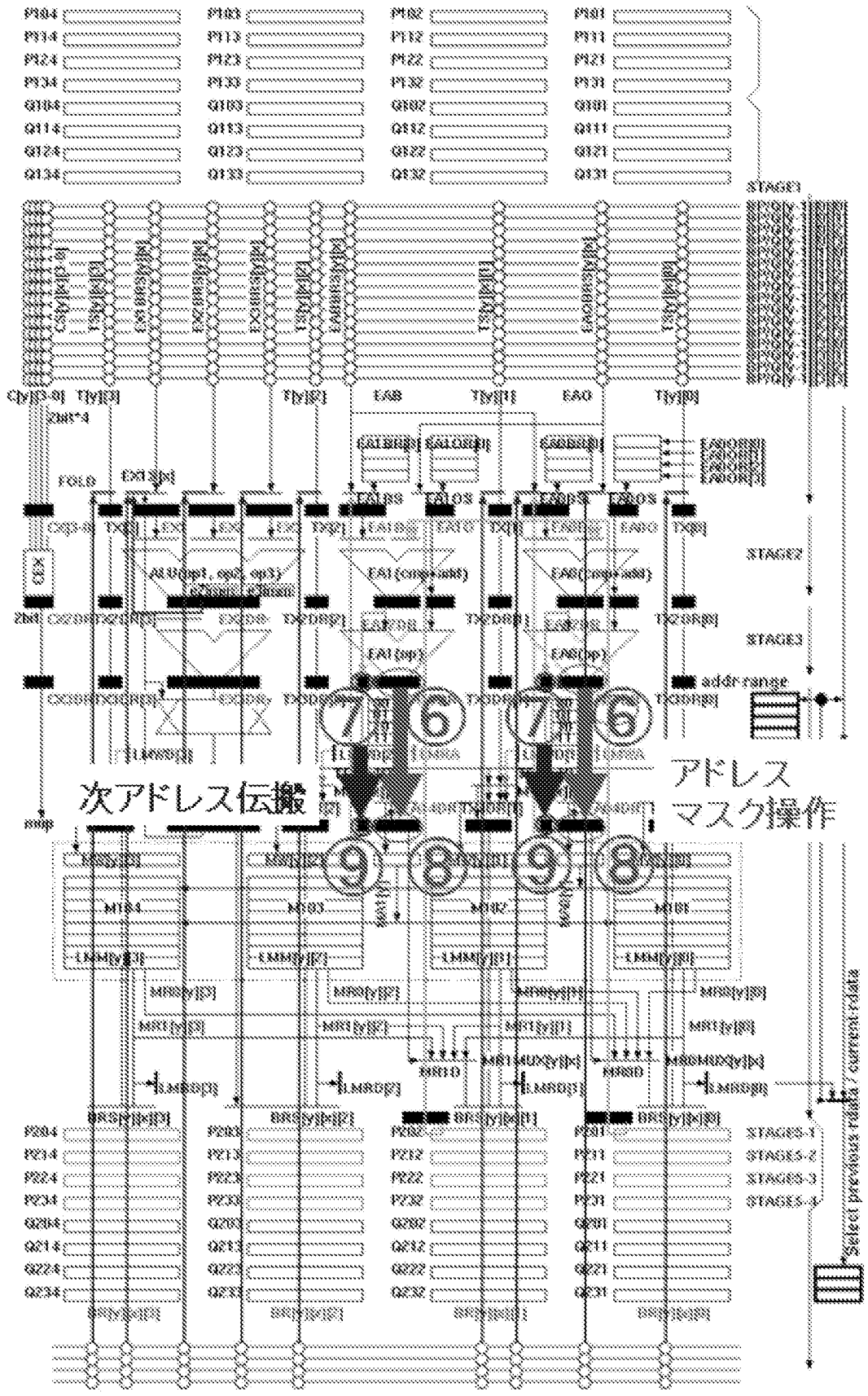
[図8]


B,A行列の
ベースアドレス
データ保持レジスタ
インデックス
値B
インデックス
値A
STAGE1
STAGE2
STAGE3
STAGE4
addr ram
Select previous rdata / current rdata

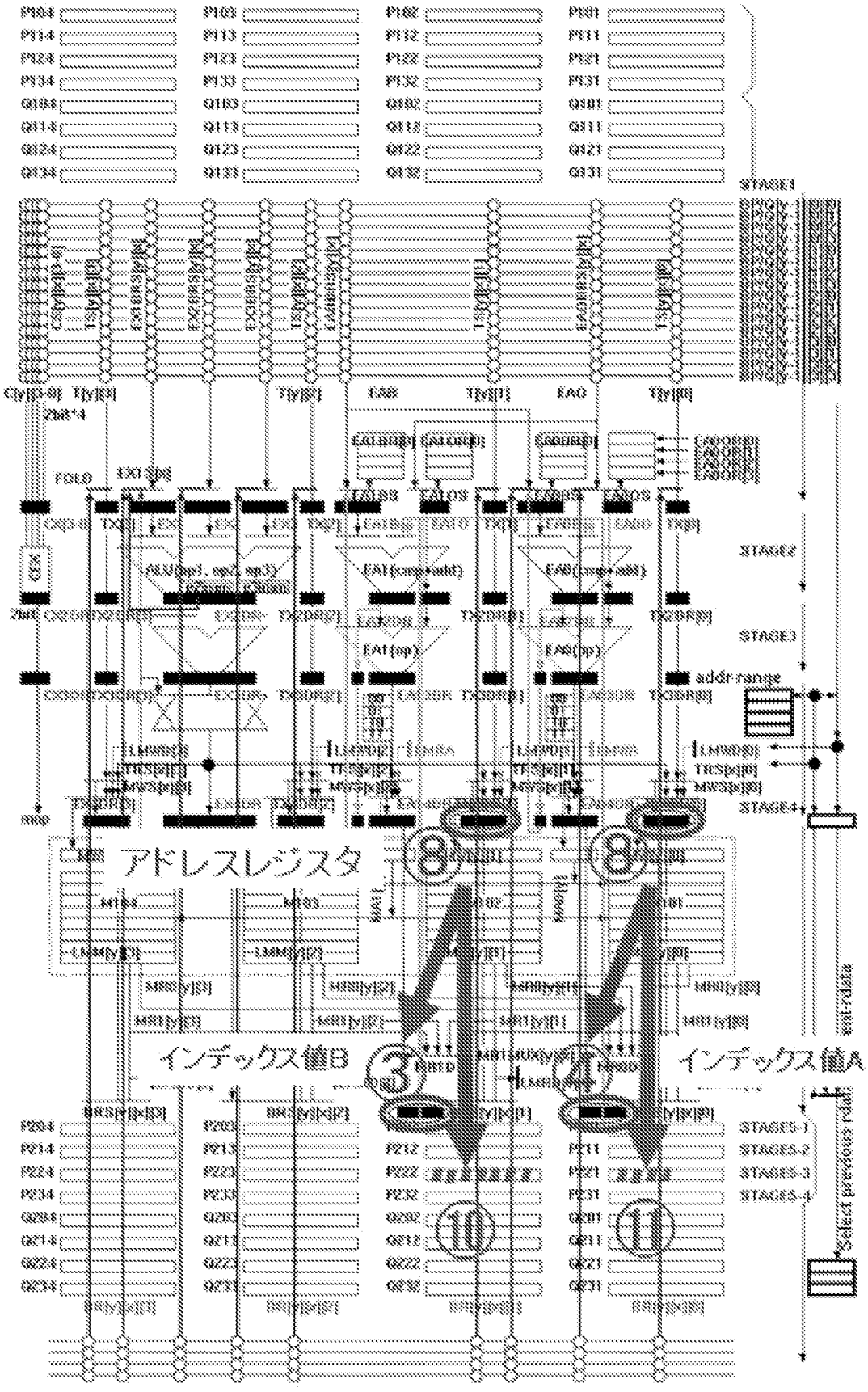
[図9]


次アドレスの保持
入力オフセット
アドレス加算
STAGE3
STAGE2
STAGE4
STAGE5-1
STAGE5-2
STAGE5-3
STAGE5-4
Select previous read / current read

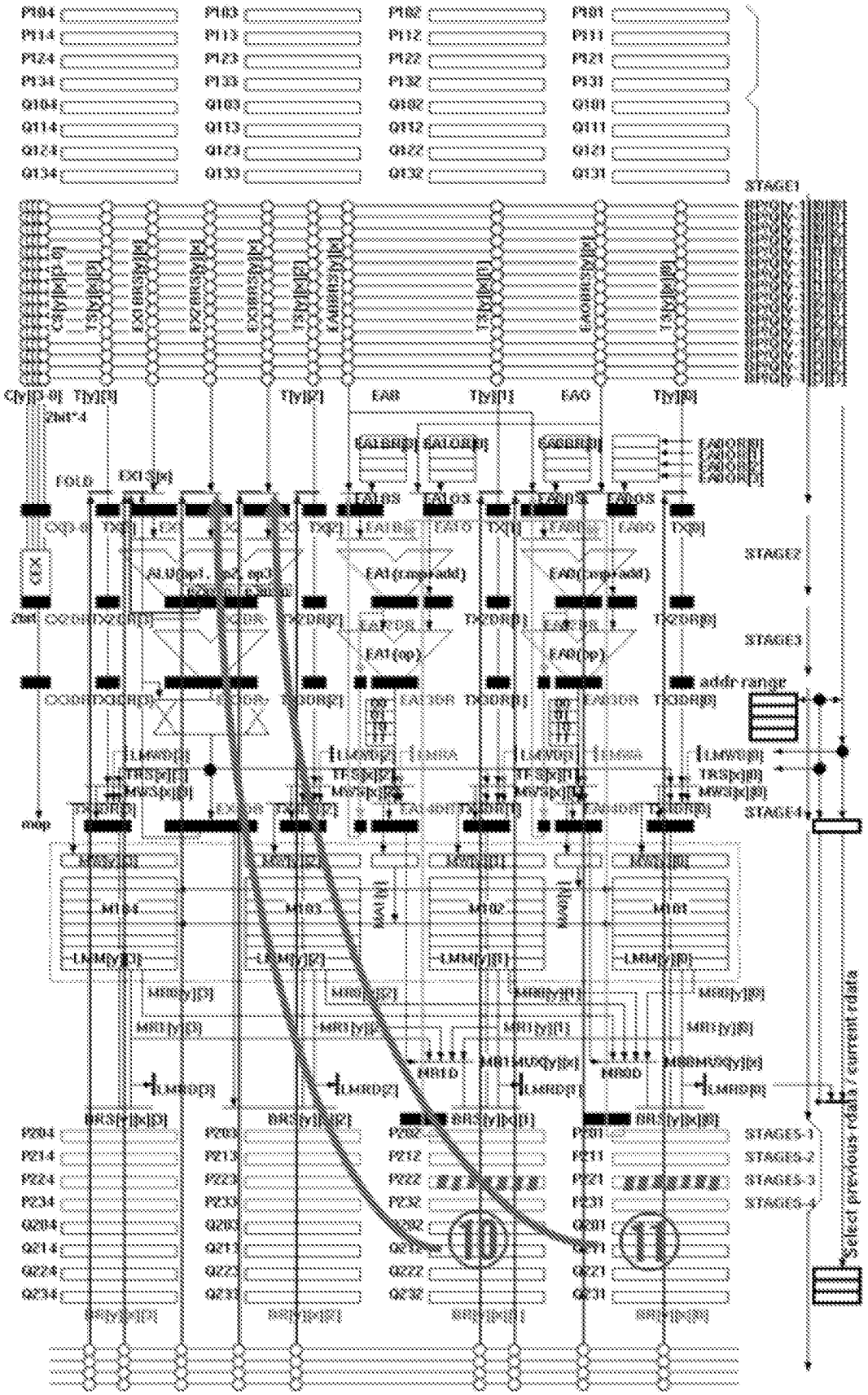
[図10]


次アドレス伝搬
アドレス
マスク操作
STAGE1
STAGE2
STAGE3
STAGE5-1
STAGE5-2
STAGE5-3
STAGE5-4
addr range
Select previous read / current data

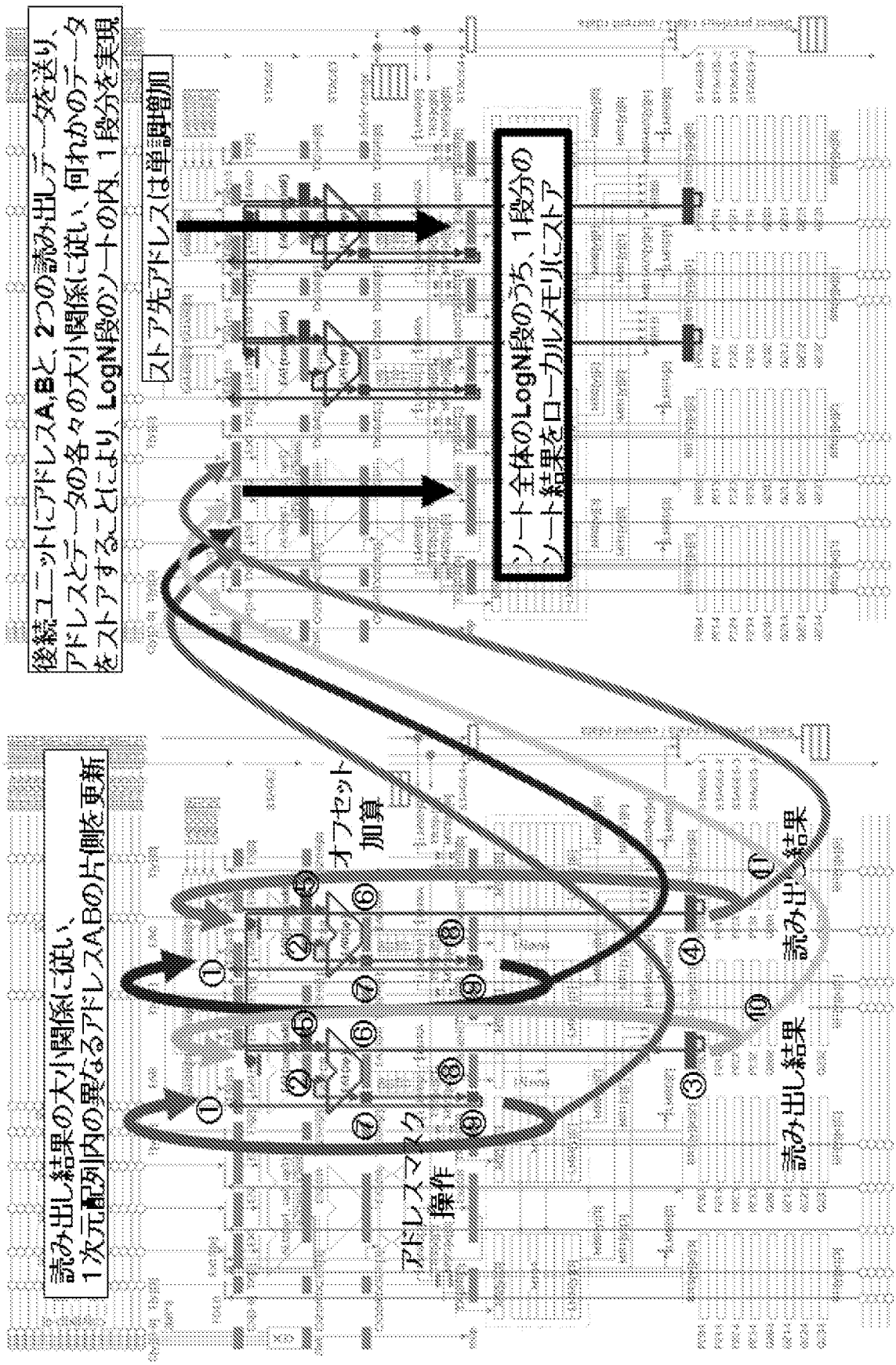
[図11]


アドレスレジスタ
インデックス値B
インデックス値A
STAGE1
STAGE2
STAGE3
STAGE4
STAGE5-1
STAGE5-2
STAGE5-3
STAGE5-4
adder range
Select previous rdata
init rdata
8
8
3
4
10
11

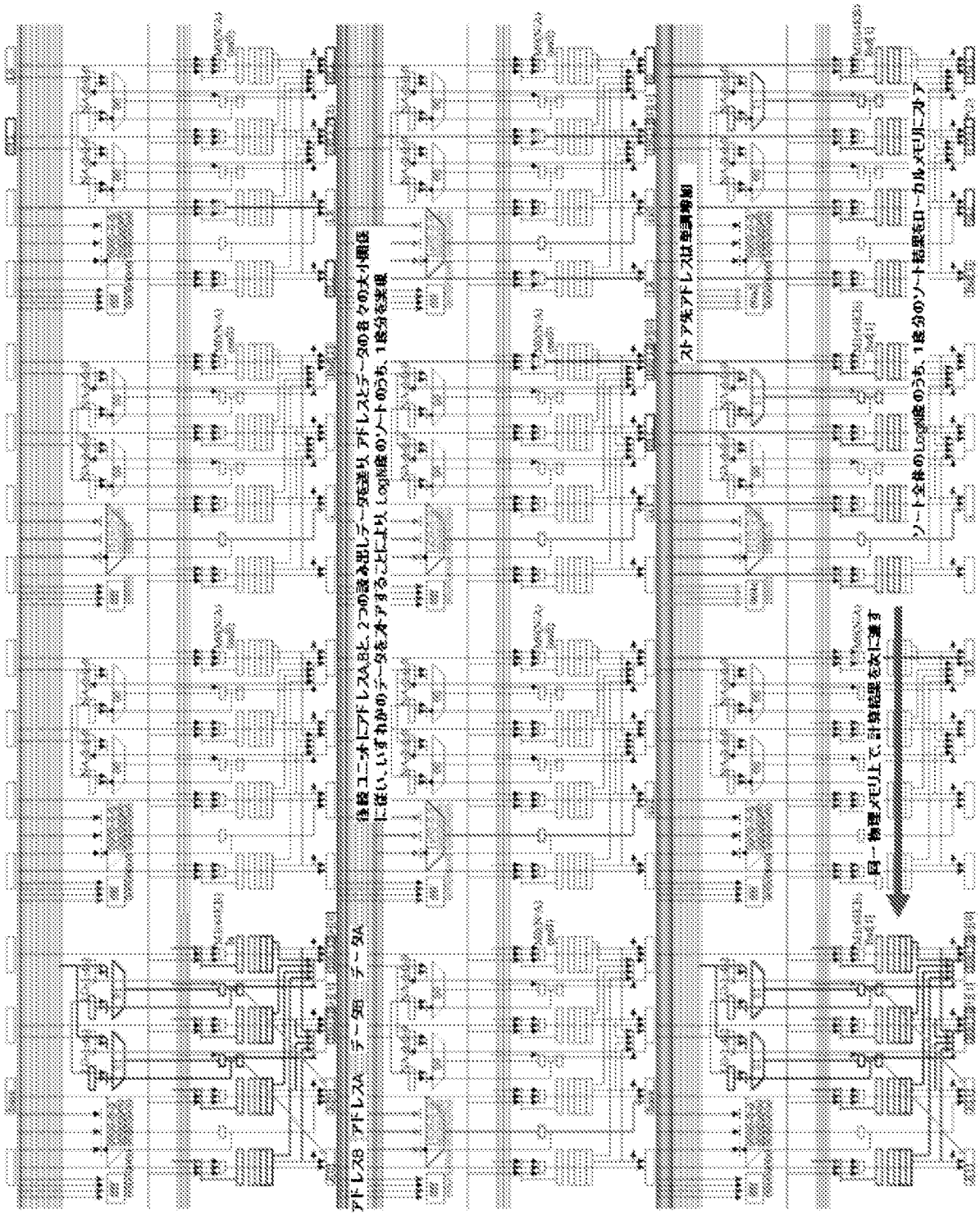
[図12]


STAGE1
STAGE2
STAGE3
STAGE4
STAGE5-1
STAGE5-2
STAGE5-3
STAGE5-4
addr range
Select previous rdata / current rdata

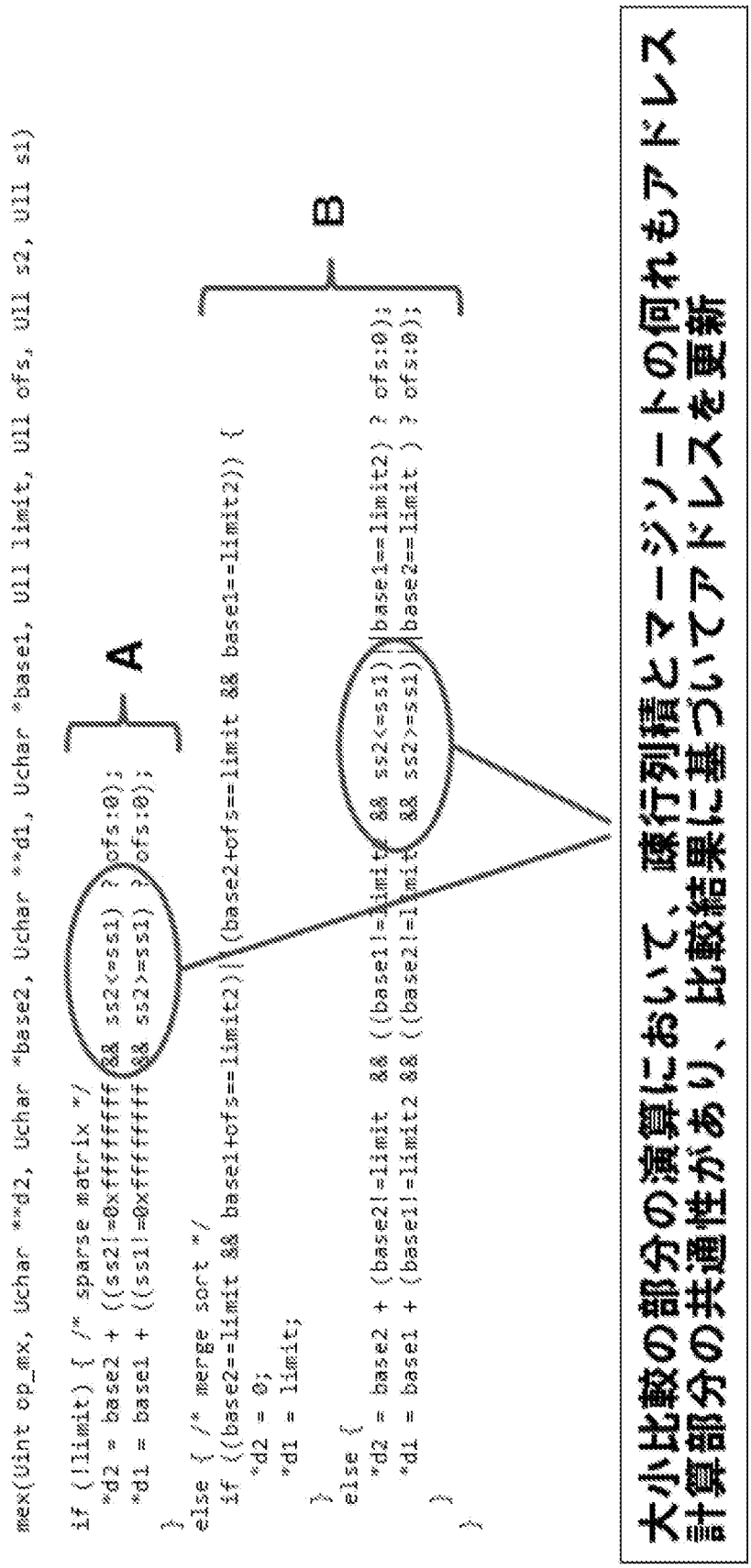
[図13]


読み出し結果の大小関係に従い、
1次元配列内の異なるアドレスABの片側を更新
後続ユニットにアドレスA,Bと、2つの読み出しデータを送り、
アドレスとデータの各々の大小関係に従い、何れかのデータ
をストアすることにより、LogN段のソートの内、1段分を実現
ストア先アドレスは単調増加
ソート全体のLogN段のうち、1段分の
ソート結果をローカルメモリにストア
アドレスマスク
操作
オフセット
加算
読み出し結果
読み出し結果

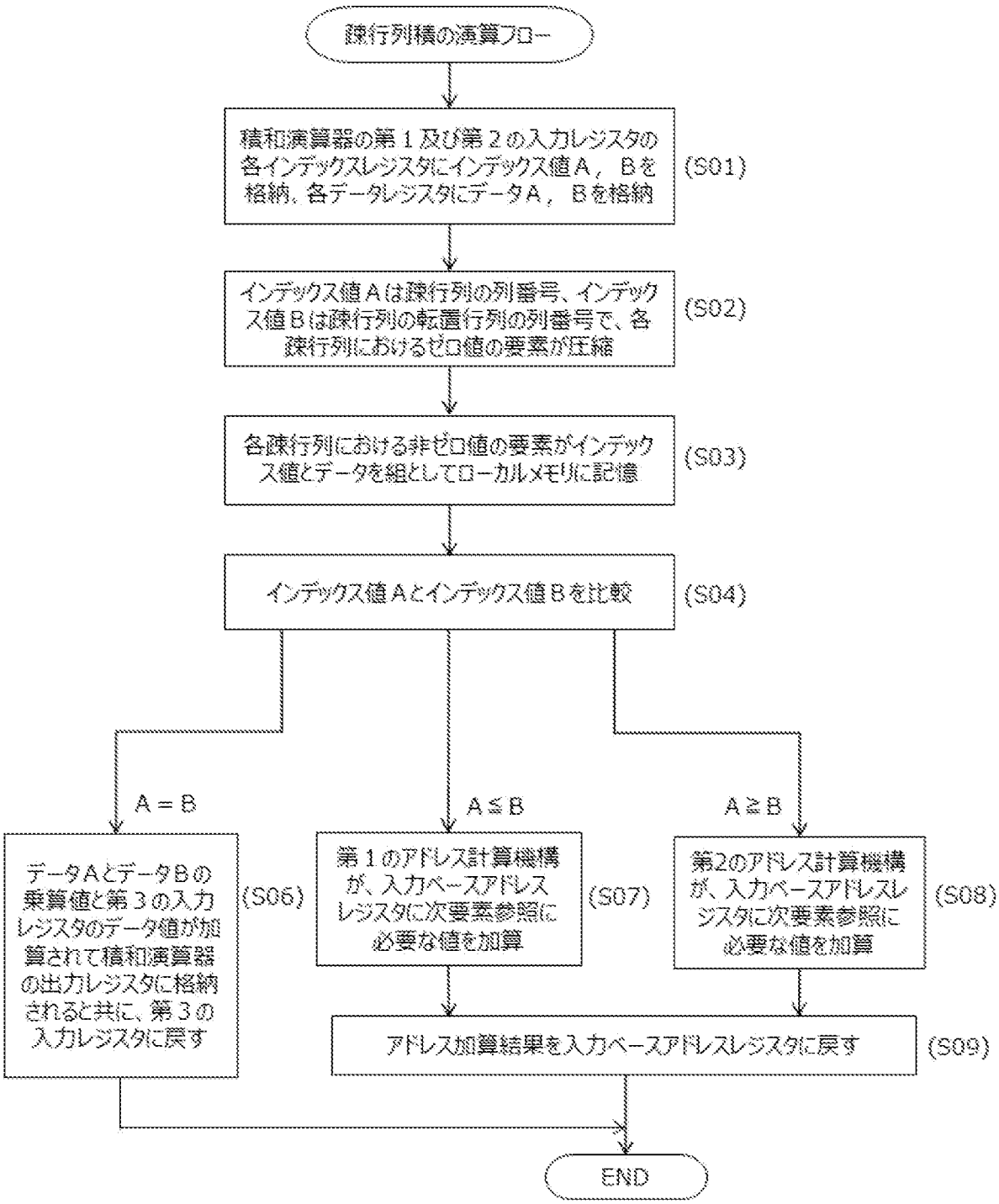
[図14]


アドレスB
アドレスA
データB
データA
後段ユニットにアドレスA,Bと、2つの読み出しデータ(Bが古い、Aが新しいデータ)を
上位、いずれかのデータを
ストア先アドレス比較情報
同一物理メモリ上で計算結果を次に渡す
ソート全体のLog2(段数)のうち、1段分のソート結果をローカルメモリにストア

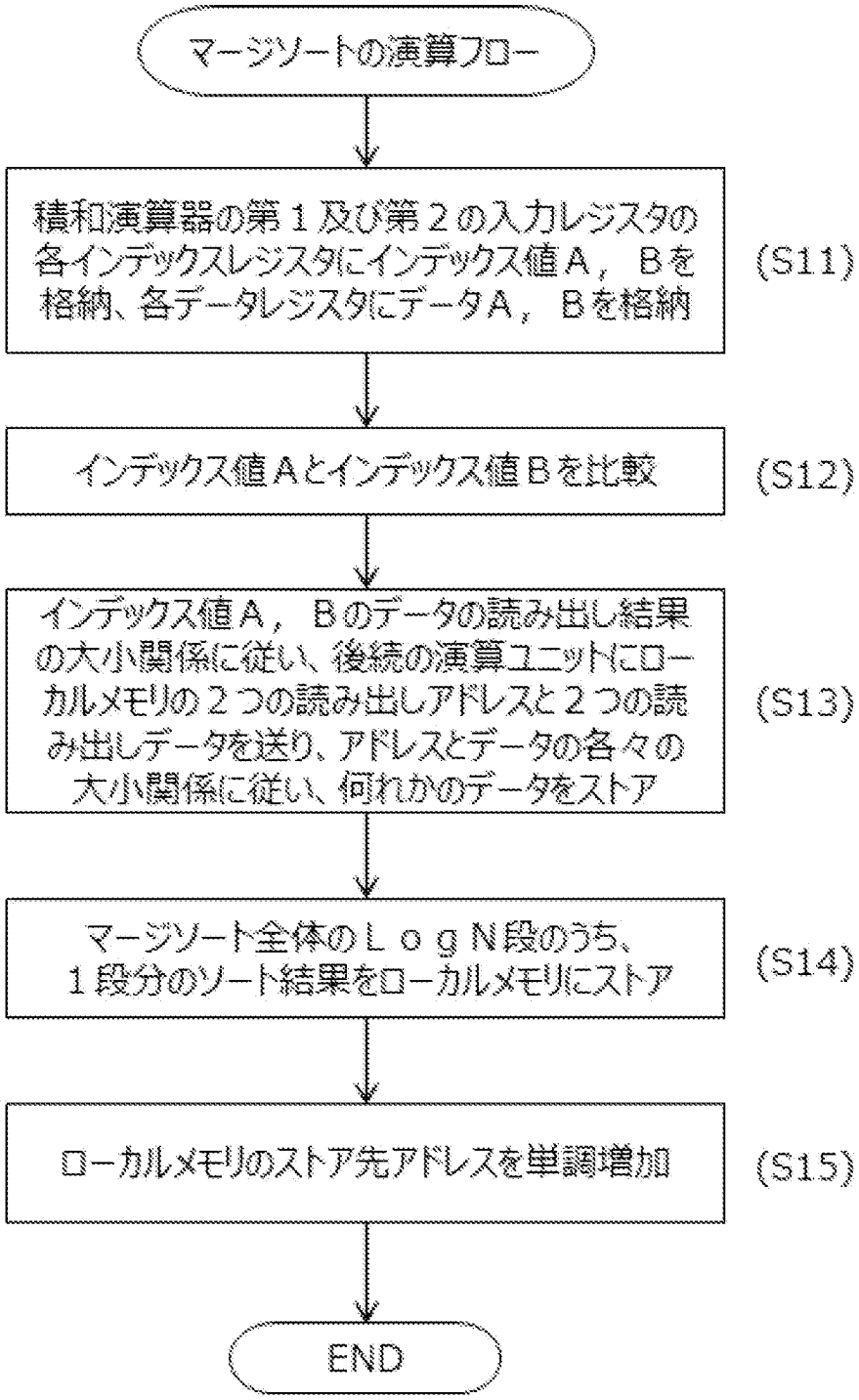
[図15]


mex(uint op_mx, Uchar **d2, Uchar *base2, Uchar **d1, Uchar *base1, Ull limit, Ull ofs, Ull s2, Ull s1)
if (!limit) { /* sparse matrix */
*d2 = base2 + ((ss2!=0xffffffffffffffff && ss2<=ss1) ? ofs:0);
*d1 = base1 + ((ss1!=0xffffffffffffffff && ss2>=ss1) ? ofs:0);
}
else { /* merge sort */
if ((base2+ofs==limit2)||(base2+ofs==limit && base1+ofs==limit2)) {
*d2 = 0;
*d1 = limit;
}
else {
*d2 = base2 + ((base2!=limit2 && ss2<=ss1)||base1==limit2) ? ofs:0);
*d1 = base1 + ((base1!=limit2 && ss2>=ss1)||base2==limit) ? ofs:0);
}
}
A
B
大小比較の部分の演算において、疎行列積とマージソートの何れもアドレス計算部分の共通性があり、比較結果に基づいてアドレスを更新

[図16]


疎行列積の演算フロー
積和演算器の第１及び第２の入力レジスタの各インデックスレジスタにインデックス値Ａ，Ｂを格納、各データレジスタにデータＡ，Ｂを格納
(S01)
インデックス値Ａは疎行列の列番号、インデックス値Ｂは疎行列の転置行列の列番号で、各疎行列におけるゼロ値の要素が圧縮
(S02)
各疎行列における非ゼロ値の要素がインデックス値とデータを組としてローカルメモリに記憶
(S03)
インデックス値Ａとインデックス値Ｂを比較
(S04)
A = B
A ≦ B
A ≧ B
データＡとデータＢの乗算値と第３の入力レジスタのデータ値が加算されて積和演算器の出力レジスタに格納されると共に、第３の入力レジスタに戻す
(S06)
第１のアドレス計算機構が、入力ベースアドレスレジスタに次要素参照に必要な値を加算
(S07)
第2のアドレス計算機構が、入力ベースアドレスレジスタに次要素参照に必要な値を加算
(S08)
アドレス加算結果を入力ベースアドレスレジスタに戻す
(S09)
END

[図17]


マージソートの演算フロー
積和演算器の第１及び第２の入力レジスタの各インデックスレジスタにインデックス値Ａ，Ｂを格納、各データレジスタにデータＡ，Ｂを格納
(S11)
インデックス値Ａとインデックス値Ｂを比較
(S12)
インデックス値Ａ，Ｂのデータの読み出し結果の大小関係に従い、後続の演算ユニットにローカルメモリの２つの読み出しアドレスと２つの読み出しデータを送り、アドレスとデータの各々の大小関係に従い、何れかのデータをストア
(S13)
マージソート全体のＬｏｇＮ段のうち、１段分のソート結果をローカルメモリにストア
(S14)
ローカルメモリのストア先アドレスを単調増加
(S15)
END

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2022/046353

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/16(2006.01)i; ***G06F 7/24***(2006.01)i; ***G06F 17/10***(2006.01)i
FI: G06F17/16 P; G06F17/10 S; G06F7/24 Z

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F17/16; G06F7/24; G06F17/10

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2023
Registered utility model specifications of Japan 1996-2023
Published registered utility model applications of Japan 1994-2023

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2019/0042538 A1 (INTEL CORPORATION) 07 February 2019 (2019-02-07) paragraphs [0043]-[0056], fig. 3-4B	1-2, 7-8
A	paragraphs [0043]-[0056], fig. 3-4B	3-6, 9-10
Y	JP 2008-234076 A (FUJITSU LIMITED) 02 October 2008 (2008-10-02) paragraphs [0006]-[0008], fig. 8	1-2, 7-8
A	paragraphs [0006]-[0008], fig. 8	3-6, 9-10

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
06 February 2023	**14 February 2023**
Name and mailing address of the ISA/JP **Japan Patent Office (ISA/JP)** **3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915** **Japan**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2022/046353

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2019/0042538 A1	07 February 2019	(Family: none)	
JP 2008-234076 A	02 October 2008	US 2008/0228846 A1 paragraphs [0012]-[0021], fig. 8	

国際調査報告

国際出願番号
PCT/JP2022/046353

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06F 17/16(2006.01)i; G06F 7/24(2006.01)i; G06F 17/10(2006.01)i
FI: G06F17/16 P; G06F17/10 S; G06F7/24 Z

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

G06F17/16; G06F7/24; G06F17/10

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０２３年
日本国実用新案登録公報	１９９６－２０２３年
日本国登録実用新案公報	１９９４－２０２３年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献のカテゴリー*	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Y	US 2019/0042538 A1 (INTEL CORPORATION) 07.02.2019 (2019-02-07) [0043]-[0056], FIGs.3-4B	1-2,7-8
A	[0043]-[0056], FIGs.3-4B	3-6,9-10
Y	JP 2008-234076 A (富士通株式会社) 02.10.2008 (2008-10-02) 【０００６】～【０００８】，【図８】	1-2,7-8
A	【０００６】～【０００８】，【図８】	3-6,9-10

☐ Ｃ欄の続きにも文献が列挙されている。

☑ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー
“A” 特に関連のある文献ではなく、一般的技術水準を示すもの
“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
“O” 口頭による開示、使用、展示等に言及する文献
“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの
“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
“Y” 特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
“&” 同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
06.02.2023	14.02.2023

名称及びあて先
日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）
田中　幸雄 5B 9191
電話番号 03-3581-1101 内線 3545

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2022/046353

引用文献	公表日	パテントファミリー文献	公表日
US 2019/0042538 A1	07.02.2019	（ファミリーなし）	
JP 2008-234076 A	02.10.2008	US 2008/0228846 A1 [0012]-[0021], FIG.8	