(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2月9日(09.02.2023)

WIPO | PCT

(10) 国際公開番号
WO 2023/013649 A1

(51) 国際特許分類:
G06F 12/126 (2016.01) *G06F 12/0875* (2016.01)
G06F 12/0866 (2016.01) *G06F 13/14* (2006.01)

(21) 国際出願番号： PCT/JP2022/029689

(22) 国際出願日： 2022年8月2日(02.08.2022)

(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 2021-129531 2021年8月6日(06.08.2021) JP

(71) 出願人: 株式会社エヌエスアイテクス (NSITEXE,INC.) [JP/JP]; 〒1080075 東京都港区港南２－１６－４ 品川グランドセントラルタワー６Ｆ Tokyo (JP). 株式会社デンソー (DENSO CORPORATION) [JP/JP]; 〒4488661 愛知県刈谷市昭和町１丁目１番地 Aichi (JP).

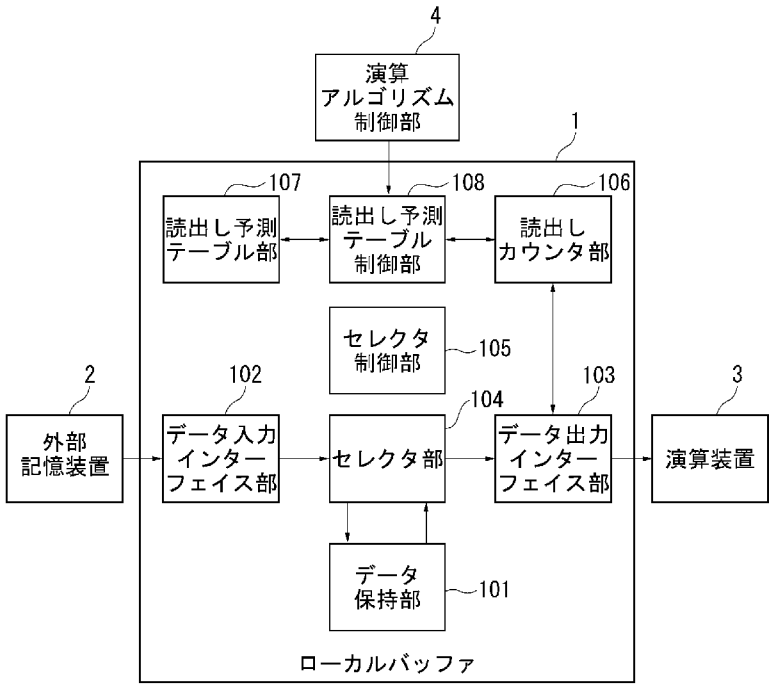
(72) 発明者: 吉田 崇 (YOSHIDA Takashi); 〒1080075 東京都港区港南２－１６－４ 品川グランドセントラルタワー６Ｆ 株式会社エヌエスアイテクス内 Tokyo (JP).

(74) 代理人: 鈴木 守, 外 (SUZUKI Mamoru et al.); 〒1500021 東京都渋谷区恵比寿西１－５－８ ＤＩＳ恵比寿ビル６階 ＫＳＩパートナーズ法律特許事務所 Tokyo (JP).

(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,

(54) **Title:** DATA CACHE DEVICE AND PROGRAM

(54) 発明の名称： データキャッシュ装置およびプログラム



(57) **Abstract:** A data cache device for temporarily holding digital data comprising: a data holding unit (101) that holds data; a data input interface unit (102) for inputting data from outside of the data cache device; a data output interface unit (103) for outputting data to outside of the data cache device; a selector unit (104) that selects one or a plurality of paths simultaneously among a path for transferring the data from the data input interface unit to the data holding unit, a path for transferring the data from the data input interface unit to the data output interface unit, and a path for transferring the

WO 2023/013649 A1

[続葉有]

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, ME, 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, 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, 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:

— 国際調査報告(条約第21条(3))

data from the data holding unit to the data output interface unit; a selector control unit (105) that controls the selector unit; and a read prediction table unit (107) that holds the number of times of planned reading of the data which is read during processing. The selector unit is controlled in accordance with the number of times of reading that is set in the read prediction table unit.

(57) 要約:データキャッシュ装置は、デジタルデータを一時的に保持するデータキャッシュ装置であって、データを保持するデータ保持部(１０１)と、前記データキャッシュ装置の外部からデータを入力するためのデータ入力インターフェイス部(１０２)と、前記データキャッシュ装置の外部へデータを出力するためのデータ出力インターフェイス部(１０３)と、前記データ入力インターフェイス部からのデータを前記データ保持部に転送する経路と、前記データ入力インターフェイス部からのデータを前記データ出力インターフェイス部に転送する経路と、前記データ保持部からのデータを前記データ出力インターフェイス部に転送する経路のうち１つまたは複数を同時に選択するセレクタ部(１０４)と、前記セレクタ部を制御するセレクタ制御部(１０５)と、前記データが処理時に読み出される予定読出し回数を保持する読出し予測テーブル部(１０７)と、を備え、前記読出し予測テーブル部に設定された読出し回数に応じて前記セレクタ部を制御する。

明　細　書

発明の名称 ： データキャッシュ装置およびプログラム

関連出願の相互参照

[0001] 本出願は、２０２１年８月６日に出願された日本出願番号２０２１－１２９５３１号に基づくものであって、その優先権の利益を主張するものであり、その特許出願の全ての内容が、参照により本明細書に組み入れられる。

技術分野

[0002] 本開示は、デジタルデータを一時的に保持するデータキャッシュ装置に関する。

背景技術

[0003] 例えば、ニューラルネットワークの処理を実行する演算装置においては、畳込み演算を連続的に実行することで推論を行う。すなわち、元の入力データに対して所定のアルゴリズムで畳込み演算を行い、その出力データに対して、さらに次のアルゴリズムで畳込み演算を実行する。

[0004] ここで、一つの入力データに対して畳込み演算を実行し出力を得る処理をレイヤと呼ぶ。ニューラルネットワークでは、各レイヤの畳込み演算のアルゴリズムが定義されており、それらレイヤの組み合わせによってネットワークを構成している。

[0005] 畳込み演算を行うためには、入力データに対してWeight値を積算し、その結果を加算する積和演算を行う。このWeight値は畳込みのアルゴリズムによってKernelサイズが規定されており、例えばKernel 3×3の処理である場合には、１つの入力データに対して縦３×横３の９つの要素のWeightを積和することになる。この畳込み処理は入力データの要素に応じて必要となるため、ニューラルネットワークの畳込み演算では大量のWeightデータの読出しが必要になる。

先行技術文献

特許文献

[0006] 特許文献1：特開２０１８－８８２５６号公報

発明の概要

[0007] このWeightを必要なタイミングで外部読出し装置から読み出すと、膨大な読出しアクセスが発生し、結果的に読出し電力が増大する。また、多くの場合、外部読出し装置からの読出し時間が処理全体のボトルネックになる可能性がある。消費電力の低減が可能な半導体装置として、例えば特許文献１が知られている。特許文献１では、キャッシュメモリ内のセットのうち、アクセス頻度が低いセットへの電源供給を遮断する。

[0008] ニューラルネットワーク処理装置では、キャッシュ装置を搭載することが考えられる。一般的なキャッシュ装置は、時間的近傍性または位置的近傍性に基づいて、一旦外部から読み出したデータを内部のキャッシュ装置に保存する。再度同じデータの読出しが発生した場合には、外部へのアクセスは行わずキャッシュ装置からデータを読み出す。このキャッシュ装置は、外部からのデータ読出しによる時間的遅延を隠ぺいする効果が期待できる。また、外部記憶装置とキャッシュ装置の実装の差異によって電力消費量を削減することが可能である。

[0009] しかしながら、ニューラルネットワーク処理においては、複数レイヤをまとめて処理することがある。すなわち、入力データを処理単位に切り出し、切り出された入力データに対して複数レイヤの処理を連続的に実行することがある。これにより、あるレイヤの出力データを即座に次の入力データとして再利用することが可能となり、不要なメモリアクセスや遅延を削減することが可能である。

[0010] 発明者の詳細な検討の結果、このような処理を行う場合には、レイヤごとのWeightを頻繁に切り替え、複雑な読出しが必要となるが、この場合には時間的近傍性や位置的近傍性が必ずしも担保されず、結果的にキャッシュミスが増大するため、従来技術のキャッシュ装置では効果的な電力削減ができないという課題が見出された。

[0011] そこで、本開示は、電力削減の効果があるデータキャッシュ装置を提供す

ることを目的とする。

[0012] 本開示は上記課題を解決するために以下の技術的手段を採用する。特許請求の範囲は、ひとつの態様として後述する実施形態に記載の具体的手段との対応関係を示す一例であって、本発明の技術的範囲を限定するものではない。

[0013] 本開示のデータキャッシュ装置は、デジタルデータを一時的に保持するデータキャッシュ装置であって、データを保持するデータ保持部と、前記データキャッシュ装置の外部からデータを入力するためのデータ入力インターフェイス部と、前記データキャッシュ装置の外部へデータを出力するためのデータ出力インターフェイス部と、前記データ入力インターフェイス部からのデータを前記データ保持部に転送する経路と、前記データ入力インターフェイス部からのデータを前記データ出力インターフェイス部に転送する経路と、前記データ保持部からのデータを前記データ出力インターフェイス部に転送する経路のうち１つまたは複数を同時に選択するセレクタ部と、前記セレクタ部を制御するセレクタ制御部と、前記データが処理時に読み出される予定読出し回数を保持する読出し予測テーブル部と、を備え、前記読出し予測テーブル部に設定された読出し回数に応じて前記セレクタ部を制御する。

[0014] 本開示のプログラムは、データを保持するデータ保持部と、前記データキャッシュ装置の外部からデータを入力するためのデータ入力インターフェイス部と、前記データキャッシュ装置の外部へデータを出力するためのデータ出力インターフェイス部と、前記データ入力インターフェイス部からのデータを前記データ保持部に転送する経路と、前記データ入力インターフェイス部からのデータを前記データ出力インターフェイス部に転送する経路と、前記データ保持部からのデータを前記データ出力インターフェイス部に転送する経路のうち１つまたは複数を同時に選択するセレクタ部と、を備えるデータキャッシュ装置を制御するためのプログラムであって、コンピュータを、前記データが処理時に読み出される予定読出し回数を保持する読出し予測テーブル手段、前記読出し予測テーブル手段に設定された読出し回数に応じて

前記セレクタ部を制御するセレクタ制御手段として機能させるためのプログラムである。

[0015] 本開示の別の態様のプログラムは、演算装置による演算を実行するのに必要なデータとして、データキャッシュ装置に保持したデータを読み出すか、外部記憶装置から読み出すかを設定するテーブルを作成するプログラムであって、コンピュータを、演算シミュレータによって前記演算装置の演算シミュレーションを行って、演算実行時の各データの読出し回数をカウントする読出しカウント手段、前記読出しカウント手段による各データの読出し回数に基づいて、前記テーブルを作成するテーブル作成手段として機能させるためのプログラムである。

図面の簡単な説明

[0016] 本開示についての上記目的およびその他の目的、特徴や利点は、添付の図面を参照しながら下記の詳細な記述により、より明確になる。その図面は、

[図1]図１は、第１の実施の形態のデータキャッシュ装置の構成を示す図であり、

[図2]図２は、第１の実施の形態のデータキャッシュ装置を含む演算装置の全体構成を示す図であり、

[図3]図３は、第１の実施の形態のデータキャッシュ装置の動作を示すフローチャートであり、

[図4]図４は、第２の実施の形態のデータキャッシュ装置の構成を示す図であり、

[図5]図５は、第３の実施の形態のデータキャッシュ装置の構成を示す図であり、

[図6]図６は、第３の実施の形態のデータキャッシュ装置の動作を示すフローチャートであり、

[図7]図７は、第４の実施の形態のデータキャッシュ装置の構成を示す図である。

発明を実施するための形態

[0017] 以下、本開示の実施の形態のデータキャッシュ装置について、図面を参照しながら説明する。

（第１の実施の形態）

図１は、第１の実施の形態のデータキャッシュ装置１の構成を示す図、図２はデータキャッシュ装置１、外部記憶装置２、演算装置３と演算アルゴリズム制御部４を含む全体構成を示す図である。図２に示すようにデータキャッシュ装置１は、外部記憶装置２と演算装置３と演算アルゴリズム制御部４に接続されている。データキャッシュ装置１は外部記憶装置２からのデータを入力し、一部のデータを保持したうえで、演算装置３に出力する。演算アルゴリズム制御部４は、ニューラルネットワークのレイヤパラメタやアルゴリズムを管理し、データキャッシュ装置１および演算装置３を連動して制御する。

[0018] 次に、本実施の形態のデータキャッシュ装置１の内部構成について説明する。１０１はデータ保持部である。このデータ保持部１０１は、本実施の形態ではニューラルネットワークのWeightデータを一時的に保持するキャッシュメモリである。該データ保持部１０１は、SRAMまたはフリップフロップ回路で実装される。なお、データ保持部１０１は、「ローカルバッファ」ともいう。

[0019] １０２はデータ入力インターフェイス部１０２である。データ入力インターフェイス部１０２は、外部記憶装置２と接続され、外部記憶装置２からのデータを受信する。データ入力インターフェイス部１０２と外部記憶装置２の接続については、中間にバスインターフェイス等があってもよく、論理的にデータが取得できる状態であればよい。データ入力インターフェイス部１０２はバスマスタとしての機能を内包し、演算装置３からセレクタ部１０４を経由して伝達された要求に応じて外部記憶装置２からデータを取得する。

[0020] １０３はデータ出力インターフェイスである。データ出力インターフェイス部１０３は演算装置３と接続されている。

[0021] １０４はセレクタ部である。セレクタ部１０４はデータ入力インターフェ

イス部１０２からの入力データの送出先とデータ出力インターフェイス部１０３の要求データの取得元を選択する回路である。

[0022] １０５はセレクタ制御部である。セレクタ制御部１０５はセレクタ部１０４を制御する部分であり、命令実行サイクル単位でセレクタ部１０４の接続を制御する。セレクタ制御部１０５は、データ入力インターフェイス部１０２からのデータをデータ保持部１０１に転送する経路と、データ入力インターフェイス部１０２からのデータをデータ出力インターフェイス部１０３に転送する経路と、データ保持部１０１からのデータをデータ出力インターフェイス部１０３に転送する経路のうち１つまたは複数を同時に選択する。

[0023] １０６は読出しカウンタ部である。読出しカウンタ部１０６は、データ出力インターフェイス部１０３のアクセス数をカウントする機能である。内部にカウンタを複数保持し、データ出力インターフェイス部１０３を経由して、演算装置３が要求するアドレスごとの読出しを独立してカウントする。

[0024] １０７は読出し予測テーブル部である。読出し予測テーブル部１０７はSRAMまたはフリップフロップ回路にて実装され、各レイヤでのWeight値が読み出される読出し回数が格納されている。予測回数はシステム起動時に初期値が書き込まれ、その後任意のタイミングで更新されてもよい。また、実装方針によっては、従前に設定された読出し回数閾値との大小関係のみを保持してもよい。

[0025] １０８は読出し予測テーブル制御部である。読出し予測テーブル制御部１０８は演算アルゴリズム制御部４から指定されたアルゴリズム情報を取得し、読出し予測テーブル部１０７から該当レイヤに対するWeightデータの読出し予測回数を読み出し、セレクタ制御部１０５に対してセレクタ選択信号を発出する。加えて、読出し予測テーブル制御部１０８は演算アルゴリズム制御部４の指示に基づいて、読出しカウンタ部１０６から読出しカウンタ情報を取得し、読出し予測テーブル部１０７の値を更新する機能を持つ。

[0026] 図３は、データキャッシュ装置１の動作を示すフローチャートである。以下に説明する例では、演算装置３は、ニューラルネットワークによる動画処

理の演算を例として説明する。演算装置３は、動画を構成するフレームごとに、画像処理を行う。より具体的には、処理対象のフレーム画像または特徴マップを複数の処理単位の領域に分割し、該領域をタイルと呼ぶが、タイルごとにニューラルネットワークによる推論処理を行う。各フレームに対して行う推論処理は同じである。

[0027] ステップＳ００１において、読出し予測テーブル制御部１０８は、演算アルゴリズム制御部４から処理するアルゴリズムのリストを受信する。例えば、ニューラルネットワークの複数レイヤを一括して処理する場合には、そのレイヤ数、各レイヤのカーネル数、入力データチャネル数および出力データチャネル数および入力アクティベーションサイズの情報などである。読出し予測テーブル制御部１０８は、受信した情報に基づいて、必要となるWeight保持量と、演算を繰り返すタイル数を算出する。または、演算アルゴリズム制御部４自身がWeight保持量を計算し、その結果を受信する実装としてもよい。

[0028] 読出し予測テーブル制御部１０８は、必要となるWeight保持量がデータ保持部の容量以下であるか否かを判定する。複数レイヤ処理分に必要となるWeight量がデータ保持部１０１の容量以下である場合には（Ｓ００１で「はい」）、読出し予測テーブル制御部１０８は、引き続きステップＳ００２の処理を実行する。

[0029] ステップＳ００２において、読出し予測テーブル制御部１０８は、複数レイヤを処理する際のタイル数の情報を演算アルゴリズム制御部４から受信し、タイル数が２以上か否かを判定する。

[0030] このとき、処理タイル数が１のときには（Ｓ００２で「いいえ」）、データをデータ保持部１０１に転送して一時的保持しておくことの電力的および性能的メリットがないため、ステップＳ００３に進み、読出し予測テーブル制御部１０８は外部記憶装置２からデータ保持部１０１を介さずに直接データを転送するように読出し予測テーブル部１０７を更新する。タイル数が２以上の場合には（Ｓ００２で「はい」）、ステップＳ００４に進み、予測テ

ーブルの全領域をデータ保持部１０１に転送する旨のフラグを立てた状態で読出し予測テーブル部１０７を更新する。なお、以上に説明したステップＳ００１とステップＳ００２の判断の順序は逆でもよい。

[0031] 続いて、ステップＳ０１２において、読出し予測テーブル部１０７にしたがって、外部記憶装置２からデータを転送する。具体的には、処理タイル数が１のときには、データ保持部１０１を介さずに直接データを転送し、処理タイル数が２以上のときには、外部記憶装置２からデータ保持部１０１へデータを転送する。

[0032] 一方、ステップＳ００１においてWeightデータの量がデータ保持部１０１の容量以上である場合には（Ｓ００１で「いいえ」）、ステップＳ００５に進み、処理タイル数が大きいレイヤのデータから順にデータ保持部１０１の容量に収まりきるまでのレイヤ分のWeightデータをデータ保持部１０１に転送するように読出し予測テーブル部１０７を更新する。

[0033] ステップＳ００６は、データ保持部１０１から読み出されるべきWeightデータを外部記憶装置２からデータ保持部１０１に対して転送するステップである。読出し予測テーブル制御部１０８はデータ保持部１０１から読出しが必要となるWeightデータについて、データ入力インターフェイス部１０２に対して読み出し要求を送出する。または、演算アルゴリズム制御部４を経由して要求を通知してもよい。

[0034] データ入力インターフェイス部１０２は新たにデータ保持部１０１から読み出されるべきWeightデータを外部記憶装置２からデータ保持部１０１に転送する。このときデータ入力インターフェイス部１０２はバスマスタとして機能し、読出し予測テーブルに保持されたWeightの位置情報から外部記憶装置２のアドレスを生成し、データ転送要求を出す。または、外部記憶装置２とデータ入力インターフェイス部１０２の間にＤＭＡコントローラなどを設けることで、同等の機能を代替してもよい。

[0035] ステップＳ００７以降は最初のタイルの処理が開始された後のステップである。ステップＳ００７は演算アルゴリズム制御部４が演算装置３に対して

処理開始の信号を送出し、演算装置３がデータ出力インターフェイス部１０３に対して必要となるWeightデータの取得要求をアドレスとともに送出するタイミングで発生する。データ出力インターフェイス部１０３は演算装置３からWeightデータの転送要求をアドレスとともに受信する。このとき、データ出力インターフェイス部１０３は、セレクタ制御部１０５に対して要求されているアドレス情報を伝達する。

[0036] セレクタ制御部１０５は読出し予測テーブル部１０７を参照し、該要求データが外部記憶装置２に存在するか、データ保持部１０１に保持されているかを判定し、セレクタ部１０４を制御して、必要な記憶域への経路を接続し、データ出力インターフェイス部１０３がデータにアクセスできるようにする。セレクタ制御部１０５は経路の確立の信号として該データが外部記憶装置２に存在するのか、データ保持部１０１に存在するのかのマップ情報をデータ出力インターフェイス部１０３に通知する。

[0037] データ出力インターフェイス部１０３は自身が要求したデータがどのレイヤのWeightデータであるかの情報と、セレクタ制御部１０５から返送されたマップ情報ごとに読出しカウンタ部１０６の要求転送データ量をインクリメントする。読出しカウンタ部１０６は、マップ情報に基づいて、Weightデータがデータ保持部からデータが読み出された読出し回数と、データ入力インターフェイス部から直接読み出された読出し回数をそれぞれにカウントし保持する。

[0038] ステップＳ００８において、演算装置３は全タイルが完了するまでデータ出力インターフェイス部１０３を経由しWeightデータを取得し続ける。この間データ出力インターフェイス部１０３はステップＳ００７を継続する。ステップＳ００８において全タイルの処理が完了した場合（Ｓ００８で「はい」）、ステップＳ００９に進む。ここまでの処理で、１つのフレームを処理したときに、各レイヤでWeightデータが読み出された回数のカウントデータが読出しカウンタ部１０６によってカウントされる。

[0039] ステップＳ００９において、読出し予測テーブル制御部１０８はレイヤご

とに読出しカウンタ部１０６からカウントデータを読み出し、演算アルゴリズム制御部４により予め指定された閾値と比較を行う。該閾値は、データの読み出し及び転送に消費される電力量から算出される。すなわち、外部記憶装置２からデータ保持部１０１を介さずにデータを転送するのに要する電力量と、外部記憶装置２からデータ保持部１０１にデータを転送し、データ保持部１０１からデータを読み出すのに要する電力量とに基づいて決定される。

[0040] 読出し回数が閾値以上の場合には、データ保持部１０１からWeightデータを読み出したほうが外部記憶装置２から直接データを取得するよりも消費電力が低くなる。読出し予測テーブル制御部１０８はWeightデータの読出し回数と閾値との大小比較の結果として、読出し回数が閾値以上のレイヤのWeightデータについてはフラグ設定する。ステップＳ０１０において、前述のステップＳ００９において設定されたフラグ情報を読み出し、読出し予測テーブル部１０７に書き込む。

[0041] なお、ステップＳ００９において閾値評価を行わず、読出しカウント数の多い順、すなわち読出し量の多いレイヤから順に読出し予測テーブル部１０７にデータ保持部１０１から読み出す旨のフラグを設定してもよい。

[0042] ステップＳ０１１は読出し予測テーブル部１０７の情報に基づいて、データ保持部１０１にデータを転送するステップである。読出し予測テーブル制御部１０８は新たにデータ保持部１０１から読出しが必要となるWeightデータについて、データ入力インターフェイス部１０２に対して読み出し要求を送出する。または、演算アルゴリズム制御部４を経由して要求を通知してもよい。データ入力インターフェイス部１０２は新たにデータ保持部１０１から読み出されるべきWeightデータを、新たに外部記憶装置２から直接読み出すようになったWeightデータの領域に上書きする。または、新たに更新された読出し予測テーブル部１０７の内容に基づき、改めてステップＳ００６を実行してもよい。

[0043] ステップＳ０１３において、上述の処理によって準備された読出し予測テ

ーブル部１０７およびデータ保持部１０１のデータを用いて、残りのフレームについての処理を実施する。

[0044] 本実施の形態においては、データ保持部１０１に保持するデータとしてWeightデータを例として説明したが、本開示はWeightデータのみに適用されるものではなく、その他のデータ、例えばアクティベーションデータやその他の重みデータに適用してもよいことは言うまでもない。

[0045] 以上、本実施の形態のデータキャッシュ装置１の構成について説明したが、セレクタ制御部１０５、読出しカウンタ部１０６、読出し予測テーブル部１０７、読出し予測テーブル制御部１０８の各機能を実現するモジュールを有するプログラムを記憶手段に格納しておき、ＣＰＵによって当該プログラムを実行することによって、上記した制御を行うデータキャッシュ装置１が実現される。このようなプログラムも本開示の範囲に含まれる。

[0046] 本実施の形態のデータキャッシュ装置１は、最初のフレーム処理を行う際に、読出し回数が閾値以上のレイヤのWeightデータにフラグを設定し、読出し予測テーブル部１０７に書き込んでおく。読出し予測テーブル部１０７に基づいてデータ保持部１０１にWeightデータを転送しておくので、残りのフレーム処理を行う際に、データ保持部１０１に記憶されたWeightデータを用いることができ、電力消費量を抑制しつつ演算性能を向上させることができる。加えて、残りのフレーム処理を行う際には、本実施の形態のカウントおよび比較処理を停止することで、評価用の電力を削減してもよい。

[0047] （第２の実施の形態）

図４は、第２の実施の形態のデータキャッシュ装置５の構成を示す図である。第２の実施の形態のデータキャッシュ装置５の基本的な構成は、第１の実施の形態のデータキャッシュ装置１と同じであるが、第２の実施の形態のデータキャッシュ装置５は、圧縮されたWeightデータを扱う点が異なる。

[0048] データキャッシュ装置５は、図２で説明した外部記憶装置２、演算装置３、及び演算アルゴリズム制御部４と接続されている。データキャッシュ装置５は外部記憶装置２からのデータを入力し、一部のデータを保持したうえで

、演算装置３に出力する。演算アルゴリズム制御部４は、ニューラルネットワークのレイヤパラメタやアルゴリズムを管理し、データキャッシュ装置５及び演算装置３を連動して制御する。

[0049] 次に、第２の実施の形態のデータキャッシュ装置５の内部構成について説明する。

５０１はデータ保持部である。このデータ保持部５０１はWeightデータを一時的に保持するキャッシュメモリである。データ保持部５０１はSRAMまたはフリップフロップ回路で実装される。

[0050] ５０２はデータ入力インターフェイス部である。データ入力インターフェイス部５０２は、外部記憶装置２と接続され、外部記憶装置２からのデータを受信する。データ入力インターフェイス部５０２と外部記憶装置２の接続については、中間にバスインターフェイス等があってもよく、論理的にデータが取得できる状態であればよい。データ入力インターフェイス部５０２はバスマスタとしての機能を内包し、演算装置３からセレクタ部５０４を経由して伝達された要求に応じて、外部記憶装置２からデータを取得する。

[0051] ５０３は、データ出力インターフェイス部である。データ出力インターフェイス部５０３は演算装置３と接続されている。

[0052] ５０４は、セレクタ部である。セレクタ部５０４はデータ入力インターフェイス部５０２からの入力データの送出先とデータ出力インターフェイス部５０３の要求データの取得元を選択する回路である。

[0053] ５０５は、セレクタ制御部である。セレクタ制御部５０５はセレクタ部５０４を制御する部分であり、命令実行サイクル単位でセレクタ部５０４の接続を制御する。

[0054] ５０６は読出しカウンタ部である。読出しカウンタ部５０６は、データ出力インターフェイス部５０３のアクセス数をカウントする機能である。内部にカウンタを複数保持し、データ出力インターフェイス部５０３を経由して、演算装置３が要求するアドレスごとの読出しを独立してカウントする。

[0055] ５０７は、読出し予測テーブル部である。読出し予測テーブル部５０７は

、SRAMまたはフリップフロップ回路にて実装され、各レイヤでのWeightデータが読み出される読出し回数が格納されている。予測回数はシステム起動時に初期値が書き込まれ、その後任意のタイミングで更新されてもよい。また、実装方針によっては、従前に設定された読出し回数閾値との大小関係のみを保持してもよい。

[0056] ５０８は、読出し予測テーブル制御部である。読出し予測テーブル制御部５０８は演算アルゴリズム制御部４から指定されたアルゴリズム情報を取得し、読出し予測テーブル部５０７から該当レイヤに対するWeightデータの読出し予測回数を読み出し、セレクタ制御部５０５に対してセレクタ選択信号を発出する。加えて、読出し予測テーブル制御部５０８は演算アルゴリズム制御部４の指示に基づいて、読出しカウンタ情報を読出しカウンタ部５０６から取得し、読出し予測テーブル部５０７の値を更新する機能を持つ。

[0057] ５０９は、データ伸長部である。データ伸長部５０９は演算装置３のデータ要求に基づいて転送されてくる圧縮されたWeightデータを所定の方法で伸長する。圧縮の方法は特定の方法に限定されることはなく、そのアルゴリズムは本開示の有効性に影響を及ぼすものではないが、たとえばランレングス符号化などであればよい。

[0058] 第２の実施の形態のデータキャッシュ装置５の動作については、図３のフローチャートに準じる。ただし、ステップＳ００２,Ｓ００３からＳ００５の予測テーブル作成においては、圧縮されたWeightのデータ量によって算出される必要がある。

[0059] 本実施の形態では、Weightデータを例に説明したが、本開示はWeightのみに適用されるものではなく、その他のデータ、例えばアクティベーションデータやその他の重みデータに適用してもよいことは言うまでもない。

[0060] 第２の実施の形態のデータキャッシュ装置５も、第１の実施の形態のデータキャッシュ装置１と同様に、電力消費量を抑制しつつ演算性能を向上させることができる。

[0061] （第３の実施の形態）

図５は、第３の実施の形態のデータキャッシュ装置６の構成を示す図である。データキャッシュ装置６は、外部記憶装置２、演算装置３、及び、演算アルゴリズム制御部４と接続されている。データキャッシュ装置６は、外部記憶装置２からのデータを入力し、一部のデータを保持したうえで、演算装置３に出力する。演算アルゴリズム制御部４は、ニューラルネットワークのレイヤパラメタやアルゴリズムを管理し、演算装置３を制御する。

[0062] ７はWeight配置プログラムである。このWeight配置プログラム７はWeightデータを外部記憶装置２に配置するか、データ保持部６０１に配置するかを決定するためのプログラムである。

[0063] ８は、演算シミュレータである。演算シミュレータ８は外部記憶装置２、演算装置３、演算アルゴリズム制御部４およびデータキャッシュ装置６の動作を論理的に模擬するソフトウェアである。

[0064] 次に、本実施の形態のデータキャッシュ装置６の内部構成について説明する。

６０１は、データ保持部である。このデータ保持部６０１はWeightデータを一時的に保持するキャッシュメモリである。データ保持部６０１はSRAMまたはフリップフロップ回路で実装される。

[0065] ６０２は、データ入力インターフェイス部である。データ入力インターフェイス部６０２は、外部記憶装置２と接続され、外部記憶装置２からのデータを受信する。データ入力インターフェイス部６０２と外部記憶装置２の接続については、中間にバスインターフェイス等があってもよく、論理的にデータが取得できる状態であればよい。データ入力インターフェイス部６０２はバスマスタとしての機能を内包し、演算装置３からセレクタ部６０４を経由して伝達された要求に応じて、外部記憶装置２からデータを取得する。

[0066] ６０３は、データ出力インターフェイス部である。データ出力インターフェイス部６０３は演算装置３と接続されている。

[0067] ６０４は、セレクタ部である。セレクタ部６０４は、データ入力インターフェイス部６０２からの入力データの送出先とデータ出力インターフェイス

部６０３の要求データの取得元を選択する回路である。セレクタ部６０４はデータ出力インターフェイス部６０３が発行する要求アドレスによってデータをデータ保持部６０１から、またはデータ入力インターフェイス部６０２への経路を接続する。

[0068] ６０５はセレクタ制御部である。セレクタ制御部６０５はデータ入力インターフェイス部６０２からデータ保持部６０１に対してデータを転送するか、データ出力インターフェイス部６０３の要求に応じてデータを送出するかの制御を行う部分である。

[0069] ６０６は、データ配置テーブルである。データ配置テーブル部６０６はSRAMまたはフリップフロップ回路にて実装され、どのレイヤのWeight値がデータ保持部６０１に保持されているかが記録されている。

[0070] 次に、Weight配置プログラム７の内部構成について説明する。

７０１は、Weight配置テーブル作成手段である。Weight配置テーブル作成手段７０１は各レイヤのWeightデータを外部記憶装置２に配置するかデータ保持部６０１に保持するかをアクセス回数予測に基づき判定する。

[0071] ７０２は読出しカウント手段であり、レイヤごとのWeight読出し回数を算出する。この読出し回数算出は演算シミュレータ８にレイヤ処理を実行させることで、演算シミュレータ８から各Weightデータの読出し回数を取得する。

[0072] なお、Weight配置テーブル作成手段７０１と読出しカウント手段７０２は、プログラムのモジュールによって構成され、Weight配置プログラム７をコンピュータにより実行することにより、その機能が発揮される。

[0073] 次に、図６は、第３の実施の形態のデータキャッシュ装置６の動作を示す図である。

Ｓ１０１は、Weight配置プログラム７が演算シミュレータ８を使用して、演算のシミュレーションを実行するステップである。Weight配置プログラム７は、演算アルゴリズム制御部４への入力情報となるアルゴリズム、カーネル数、レイヤ数、入力データサイズ等の情報に基づいて、演算シミュレータ

８を用いてWeightデータへのアクセス回数を算出する。該処理をステップＳ１０２において、Weight配置テーブル作成手段７０１は全タイル・レイヤ分実行し、レイヤごとのアクセス数を読出しカウント手段７０２にリスト化する。ステップＳ１０２が完了後（ステップＳ１０２で「はい」）、ステップＳ１０３に進む。

[0074] ステップＳ１０３は、レイヤごとのWeightアクセス回数を集計するステップである。ステップＳ１０３においてWeight配置テーブル作成手段７０１は読出しカウント手段７０２に保持されたレイヤごとの読出しカウントを降順にソートする。

[0075] ステップＳ１０４からステップＳ１０６はWeight配置テーブルを作成するステップである。Weight配置テーブル作成手段７０１は、従前に取得したデータ保持部６０１の容量に基づき、Weightアクセス数の多い順にデータ保持部６０１にデータを配置するように配置テーブルのアドレスを設定する。例えば、データ保持部６０１のアドレスと外部記憶装置２のアドレスが独立に設計されている場合、データ保持部６０１のアドレスを該アクセス数の多いWeight領域にマッピングするようにすればよい。ステップＳ１０４の処理はアクセス数の多い順に順次実施し、ステップＳ１０５で、データ保持部６０１のデータ保持量がいっぱいになったかを判定する。ステップＳ１０５でデータ保持部６０１の領域を使い切った場合には（ステップＳ１０５で「はい」）、ステップＳ１０６に進む。

[0076] ステップＳ１０６は、外部記憶装置２に配置する領域を指定するステップである。当該ステップＳ１０６において、データ保持部６０１に保持しきれなかったレイヤのWeightデータに対して外部記憶装置２のアドレスを指定する。

[0077] ステップＳ１０７は、作成されたWeight配置テーブルを、データキャッシュ装置６のデータ配置テーブル部６０６に転送するステップである。該ステップＳ１０７では、実際の処理を実施するためにWeight配置プログラム７は作成したWeight配置テーブル情報をデータ配置テーブル部６０６に転送する

。転送手段はデータ入力インターフェイス部１０２経由で読み込んでもよいし、データ配置テーブル部６０６を外部領域からメモリマップしておいて、不図示の外部マスタなどを用いて転送してもよい。

[0078] ステップＳ１０８は転送されたデータ配置テーブル部６０６に準じてデータ保持部６０１から読み出されるべきWeightデータをデータ保持部６０１に転送するステップである。データ入力インターフェイス部６０２はデータ配置テーブルを読出し、セレクタ制御部６０５によってセレクタ部６０４をデータ保持部６０１に切り替え、データ保持部６０１のアドレスが記載されているWeightを外部記憶装置２から転送する。該転送は、データ入力インターフェイス部１０２を用いて実施してもよいし、外部のマスタをデータ保持部６０１に接続して転送してもよい。

[0079] ステップＳ１０９は演算処理中にデータを転送するステップである。このステップにおいて、演算装置３が要求されるデータは、データ出力インターフェイス部６０３経由でセレクタ部６０４に到達し、データ保持部６０１またはデータ入力インターフェイス部６０２から取得される。データ出力インターフェイス部６０３は演算装置３から必要となるWeightデータを受信し、データ配置テーブルを参照して、当該Weightが格納されている領域へのアドレスを生成する。演算処理中は該ステップＳ１０９を処理が完了するまで継続する。

[0080] 第３の実施の形態のデータキャッシュ装置６は、Weight配置プログラム７と演算シミュレータ８を備え、Weightデータを外部記憶装置２に配置するかデータ保持部６０１に保持するかを演算シミュレーション結果に基づいて判定し、判定結果に基づいてデータ配置テーブル部６０６を更新する。このデータ配置テーブル部６０６に従ってWeightデータを配置することで、電力消費量を抑制しつつ演算性能を向上させることができる。

[0081] （第４の実施の形態）

図７は、第４の実施の形態のデータキャッシュ装置９の構成を示す図である。第４の実施の形態のデータキャッシュ装置９の基本的な構成は、第１の

実施の形態のデータキャッシュ装置１と同じであるが、第４の実施の形態のデータキャッシュ装置９は、出力先の演算装置３を３台備えている点が異なる。なお、図７では、出力先が３台の場合を例示したが、出力先の演算装置３の台数は３台に限らず、２台あるいは４台以上であってもよい。

[0082] データキャッシュ装置９は、図２で説明した外部記憶装置２、演算装置３、及び演算アルゴリズム制御部４と接続されている。データキャッシュ装置９は外部記憶装置２からのデータを入力し、一部のデータを保持したうえで、演算装置３に出力する。演算アルゴリズム制御部４は、ニューラルネットワークのレイヤパラメタやアルゴリズムを管理し、データキャッシュ装置９及び演算装置３を連動して制御する。

[0083] 次に、第４の実施の形態のデータキャッシュ装置９の内部構成について説明する。

９０１はデータ保持部である。このデータ保持部９０１はWeightデータを一時的に保持するキャッシュメモリである。データ保持部９０１はSRAMまたはフリップフロップ回路で実装される。

[0084] ９０２はデータ入力インターフェイス部である。データ入力インターフェイス部９０２は、外部記憶装置２と接続され、外部記憶装置２からのデータを受信する。データ入力インターフェイス部９０２と外部記憶装置２の接続については、中間にバスインターフェイス等があってもよく、論理的にデータが取得できる状態であればよい。データ入力インターフェイス部９０２はバスマスタとしての機能を内包し、演算装置３からセレクタ部９０４を経由して伝達された要求に応じて、外部記憶装置２からデータを取得する。

[0085] ９０３は、データ出力インターフェイス部である。データ出力インターフェイス部９０３は３台の演算装置３と接続されている。

[0086] ９０４は、セレクタ部である。セレクタ部９０４はデータ入力インターフェイス部９０２からの入力データの送出先とデータ出力インターフェイス部９０３の要求データの取得元を選択する回路である。

[0087] ９０５は、セレクタ制御部である。セレクタ制御部９０５はセレクタ部９

０４を制御する部分であり、命令実行サイクル単位でセレクタ部９０４の接続を制御する。

[0088] ９０６は読出しカウンタ部である。読出しカウンタ部９０６は、データ出力インターフェイス部９０３のアクセス数をカウントする機能である。内部にカウンタを複数保持し、データ出力インターフェイス部９０３を経由して、それぞれの演算装置３について演算装置３が要求するアドレスごとの読出しを独立してカウントする。

[0089] ９０７は、読出し予測テーブル部である。読出し予測テーブル部９０７は、SRAMまたはフリップフロップ回路にて実装され、各レイヤでのWeight値が読み出される読出し回数が格納されている。本実施の形態では、読出し回数は演算装置３ごとに格納される。予測回数はシステム起動時に初期値が書き込まれ、その後任意のタイミングで更新されてもよい。また、実装方針によっては、従前に設定された読出し回数閾値との大小関係のみを保持してもよい。

[0090] ９０８は、読出し予測テーブル制御部である。読出し予測テーブル制御部９０８は演算アルゴリズム制御部４から指定されたアルゴリズム情報を取得し、読出し予測テーブル部９０７から該当レイヤに対するWeightデータの読出し予測回数を読み出し、セレクタ制御部９０５に対してセレクタ選択信号を発出する。加えて、読出し予測テーブル制御部９０８は演算アルゴリズム制御部４の指示に基づいて、読出しカウンタ情報を読出しカウンタ部９０６から取得し、読出し予測テーブル部９０７の値を更新する機能を持つ。

[0091] 第４の実施の形態のデータキャッシュ装置９の動作については、図３のフローチャートに準じる。ただし、ステップＳ００７において、データ出力インターフェイス部１０３が自身が要求したデータが、どの演算装置からの要求であるかの情報と、どのレイヤのWeightであるかの情報と、セレクタ制御部１０５から返送されたマップ情報ごとに読出しカウンタ部１０６の要求転送データ量をインクリメントする。

[0092] また、Ｓ００９において、データの読出し回数と閾値とを比較する際に用

いるデータの読出し回数は、処理に用いる演算装置３に紐付けられた読出し回数を用いる。例えば、読出し予測テーブル部９０７に３台の演算装置Ａ，Ｂ，Ｃのそれぞれに関連付けて、各レイヤのWeightデータの読出し回数が記憶されているとする。これから行う処理で、演算装置Ａ，Ｂを用い、演算装置Ｃを用いないとすると、演算装置Ａ，Ｂに紐づけられた読出し回数の合計値と閾値とを比較し、読出し回数が閾値以上のWeightデータにフラグを立てる。

[0093] 本実施の形態では、Weightデータを例に説明したが、本開示はWeightのみに適用されるものではなく、その他のデータ、例えばアクティベーションデータやその他の重みデータに適用してもよいことは言うまでもない。

[0094] 第４の実施の形態のデータキャッシュ装置９も、第１の実施の形態のデータキャッシュ装置１と同様に、電力消費量を抑制しつつ演算性能を向上させることができる。

[0095] また、第４の実施の形態のデータキャッシュ装置９は、演算装置３ごとに読出し回数をカウントして、読出し予測テーブル部９０７に読出し回数を記憶しておくので、これから行う処理を実行する演算装置３に合わせて、データ保持部９０１に記憶するWeightデータを決定することができる。

[0096] 以上、本開示の実施の形態に係るデータキャッシュ装置について説明したが、本開示は上記した実施の形態に限定されるものではない。

上記した実施の形態では、全タイルの処理が完了したところで（図３，Ｓ００８で「はい」）、データの読出し回数が閾値以上のデータにフラグを立て（Ｓ００９）、読出し予測テーブルを更新し（Ｓ０１０）、更新後の読出し予測テーブルに基づいてデータ保持部にデータを転送する（Ｓ０１１）例を説明した。つまり、データ保持部１０１にデータ保持しておくと電力消費量の低減に効果のあるデータをフレーム単位で決定しているが、この単位は必ずしもフレーム単位でなくてもよく、例えば、タイルの単位でもよい。すなわち、１タイルの処理が終了したところで、読出し回数が閾値以上のレイヤのデータのフラグを立ててもよい。特に、レイヤのかかわらずタイルのサ

イズがニューラルネットワークの場合には、この構成が有効である。

請求の範囲

[請求項1] デジタルデータを一時的に保持するデータキャッシュ装置であって、

データを保持するデータ保持部（１０１）と、

前記データキャッシュ装置の外部からデータを入力するためのデータ入力インターフェイス部（１０２）と、

前記データキャッシュ装置の外部へデータを出力するためのデータ出力インターフェイス部（１０３）と、

前記データ入力インターフェイス部からのデータを前記データ保持部に転送する経路と、前記データ入力インターフェイス部からのデータを前記データ出力インターフェイス部に転送する経路と、前記データ保持部からのデータを前記データ出力インターフェイス部に転送する経路のうち１つまたは複数を同時に選択するセレクタ部（１０４）と、

前記セレクタ部を制御するセレクタ制御部（１０５）と、

前記データが処理時に読み出される予定読出し回数を保持する読出し予測テーブル部（１０７）と、を備え、

前記読出し予測テーブル部に設定された読出し回数に応じて前記セレクタ部を制御するデータキャッシュ装置。

[請求項2] 請求項１に記載のデータキャッシュ装置であって、

前記データ出力インターフェイス部からの読出し回数をカウントする読出しカウンタ部（１０６）と、

前記読出し予測テーブル部にアルゴリズムに応じた読出し回数を設定する読出し予測テーブル制御部（１０８）と、をさらに備え、

前記読出し予測テーブル制御部は、前記読出しカウンタ部の出力に応じて、前記読出し予測テーブルを更新するデータキャッシュ装置。

[請求項3] 請求項２に記載されたデータキャッシュ装置であって、

前記読出しカウンタ部は、前記データ保持部からデータが読み出さ

れた読出し回数と、前記データ入力インターフェイス部から直接読み出された読出し回数をそれぞれにカウントし保持するデータキャッシュ装置。

[請求項4] 請求項２に記載されたデータキャッシュ装置であって、

前記読出し予測テーブル制御部は、前記データキャッシュ装置からデータを取得して処理を行う演算装置（３）が実行するアルゴリズムのリストを具備し、アルゴリズムに応じた予測読出し回数を計算し、読出し予測テーブル部を更新するデータキャッシュ装置。

[請求項5] 請求項２に記載されたデータキャッシュ装置であって、

前記読出しカウンタ部は、前記データキャッシュ装置がデータを出力する出力先ごとに読出し回数をカウントし、

前記読出し予測テーブル制御部は、前記出力先ごとの前記読出し回数により前記読出し予測テーブルを更新するキャッシュ装置。

[請求項6] データを保持するデータ保持部と、データキャッシュ装置の外部からデータを入力するためのデータ入力インターフェイス部と、前記データキャッシュ装置の外部へデータを出力するためのデータ出力インターフェイス部と、前記データ入力インターフェイス部からのデータを前記データ保持部に転送する経路と、前記データ入力インターフェイス部からのデータを前記データ出力インターフェイス部に転送する経路と、前記データ保持部からのデータを前記データ出力インターフェイス部に転送する経路のうち１つまたは複数を同時に選択するセレクタ部と、を備えるデータキャッシュ装置を制御するためのプログラムであって、コンピュータを、

前記データが処理時に読み出される予定読出し回数を保持する読出し予測テーブル手段、

前記読出し予測テーブル手段に設定された読出し回数に応じて前記セレクタ部を制御するセレクタ制御手段、

として機能させるためのプログラム。

[請求項7] 請求項６に記載のプログラムであって、コンピュータを、

前記データ出力インターフェイス部からの読出し回数をカウントする読出しカウンタ手段、

前記読出し予測テーブル手段にアルゴリズムに応じた読出し回数を設定すると共に、前記読出しカウンタ手段の出力に応じて前記読出し予測テーブルを更新する読出し予測テーブル制御手段、

として機能させるためのプログラム。

[請求項8] 請求項７に記載されたプログラムであって、

前記読出しカウンタ手段を、前記データ保持部からデータが読み出された読出し回数と、前記データ入力インターフェイス部から直接読み出された読出し回数をそれぞれにカウントし保持するようにコンピュータを機能させるためのプログラム。

[請求項9] 請求項７に記載されたプログラムであって、

前記読出し予測テーブル制御手段を、前記データキャッシュ装置からデータを取得して処理を行う演算装置が実行するアルゴリズムのリストを具備し、アルゴリズムに応じた予測読出し回数を計算し、読出し予測テーブル手段を更新するようにコンピュータを機能させるためのプログラム。

[請求項10] 請求項７に記載されたプログラムであって、

前記読出しカウンタ手段を、前記データキャッシュ装置がデータを出力する出力先ごとに読出し回数をカウントさせ、

前記読出し予測テーブル制御手段を、前記出力先ごとの前記読出し回数により前記読出し予測テーブル手段を更新するようにコンピュータを機能させるためのプログラム。

[請求項11] 演算装置による演算を実行するのに必要なデータとして、データキャッシュ装置に保持したデータを読み出すか、外部記憶装置から読み出すかを設定するテーブルを作成するプログラムであって、コンピュータを、

演算シミュレータによって前記演算装置の演算シミュレーションを行って、演算実行時の各データの読出し回数をカウントする読出しカウント手段（７０２）、

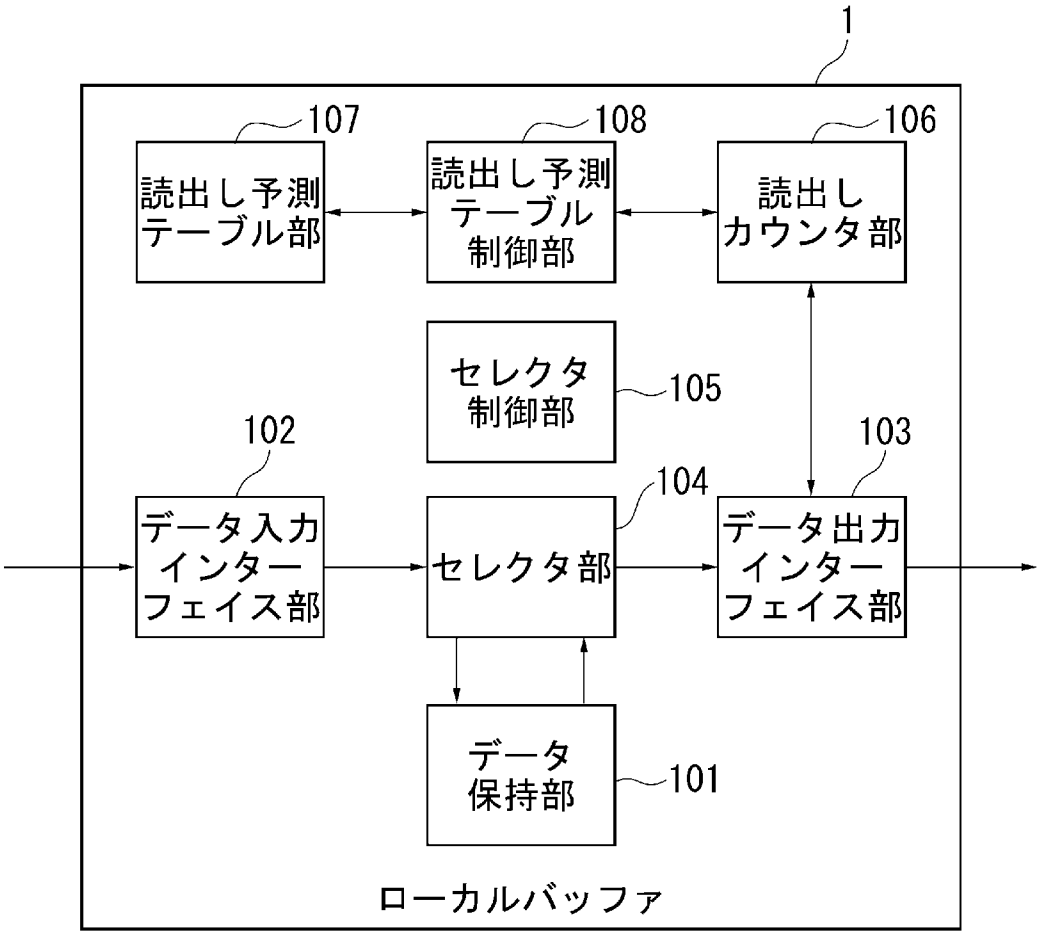
前記読出しカウント手段による各データの読出し回数に基づいて、前記テーブルを作成するテーブル作成手段（７０１）、

として機能させるためのプログラム。

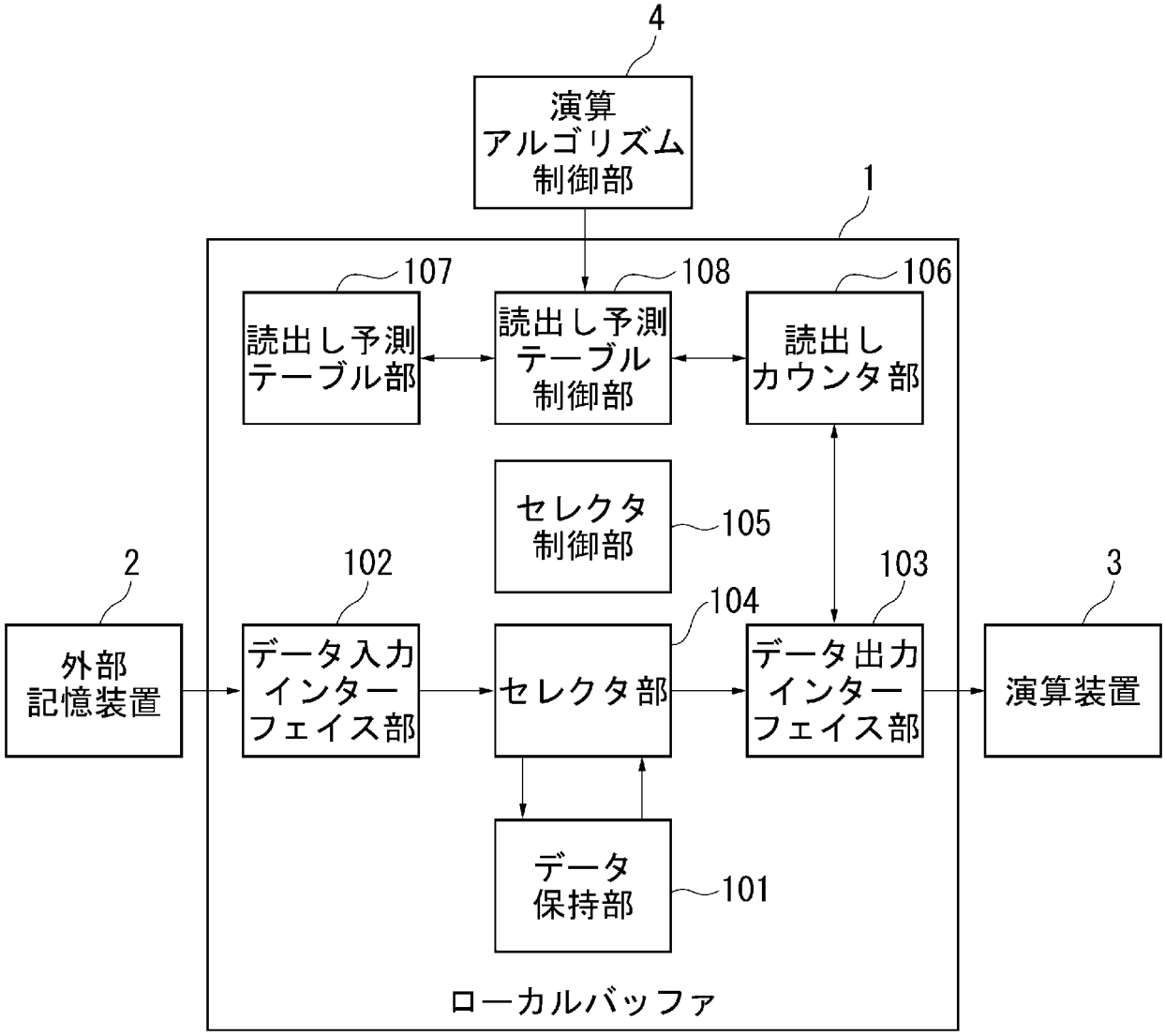
[請求項12] 請求項１１に記載のプログラムであって、

前記テーブル作成手段が、前記各データの読出し回数と前記データキャッシュ装置のデータ保持部のサイズとに基づいて前記テーブルを作成するようにコンピュータを機能させるためのプログラム。

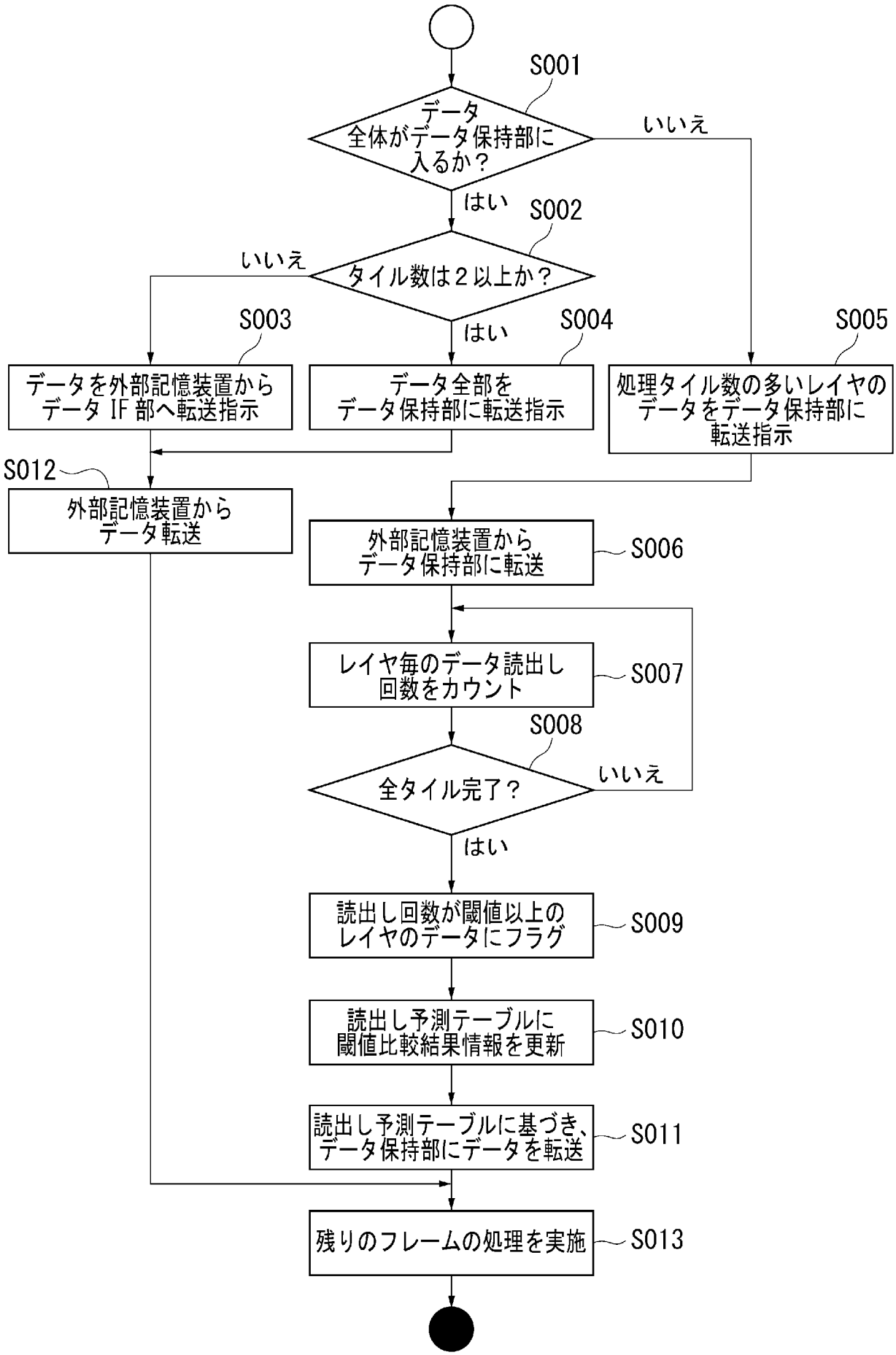
[図1]


1
107
108
106
読出し予測テーブル部
読出し予測テーブル制御部
読出しカウンタ部
セレクタ制御部
105
102
103
104
データ入力インターフェイス部
セレクタ部
データ出力インターフェイス部
データ保持部
101
ローカルバッファ

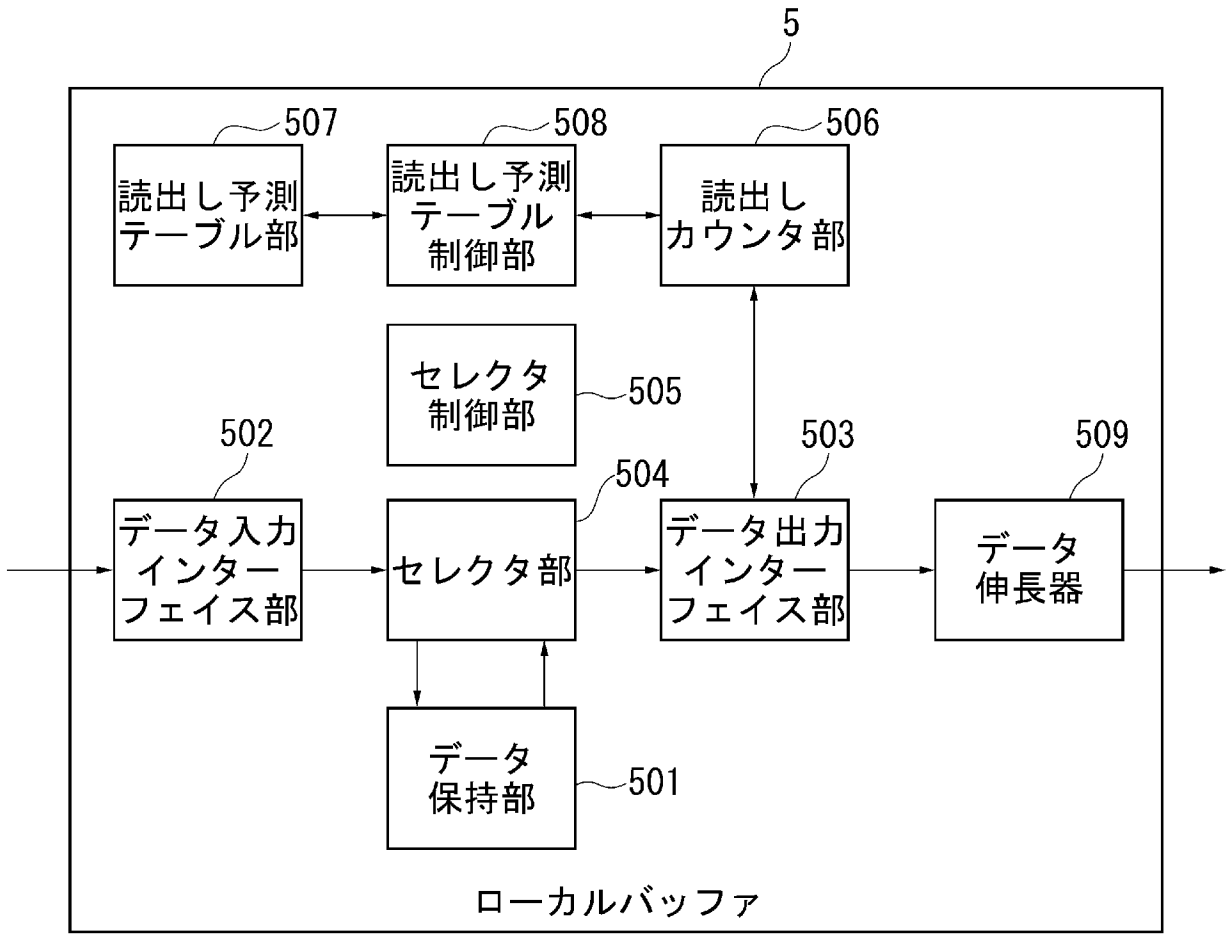
[図2]


4
演算
アルゴリズム
制御部
1
107
108
106
読出し予測
テーブル部
読出し予測
テーブル
制御部
読出し
カウンタ部
セレクタ
制御部
105
2
102
103
3
104
外部
記憶装置
データ入力
インター
フェイス部
セレクタ部
データ出力
インター
フェイス部
演算装置
データ
保持部
101
ローカルバッファ

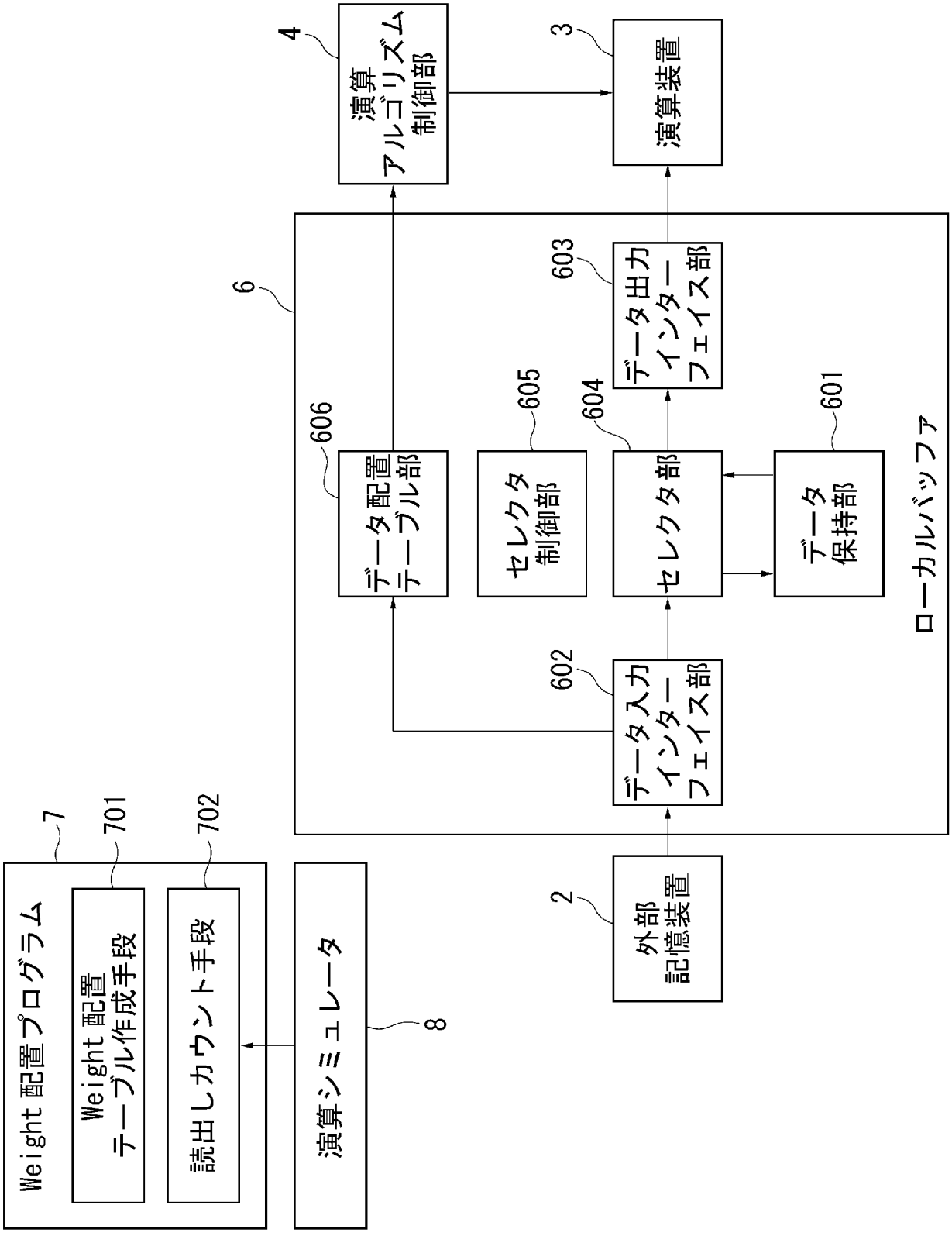
[図3]


S001
データ全体がデータ保持部に入るか？
いいえ
はい
S002
タイル数は２以上か？
いいえ
はい
S003
データを外部記憶装置からデータ IF 部へ転送指示
S004
データ全部をデータ保持部に転送指示
S005
処理タイル数の多いレイヤのデータをデータ保持部に転送指示
S012
外部記憶装置からデータ転送
外部記憶装置からデータ保持部に転送
S006
レイヤ毎のデータ読出し回数をカウント
S007
S008
全タイル完了？
いいえ
はい
読出し回数が閾値以上のレイヤのデータにフラグ
S009
読出し予測テーブルに閾値比較結果情報を更新
S010
読出し予測テーブルに基づき、データ保持部にデータを転送
S011
残りのフレームの処理を実施
S013

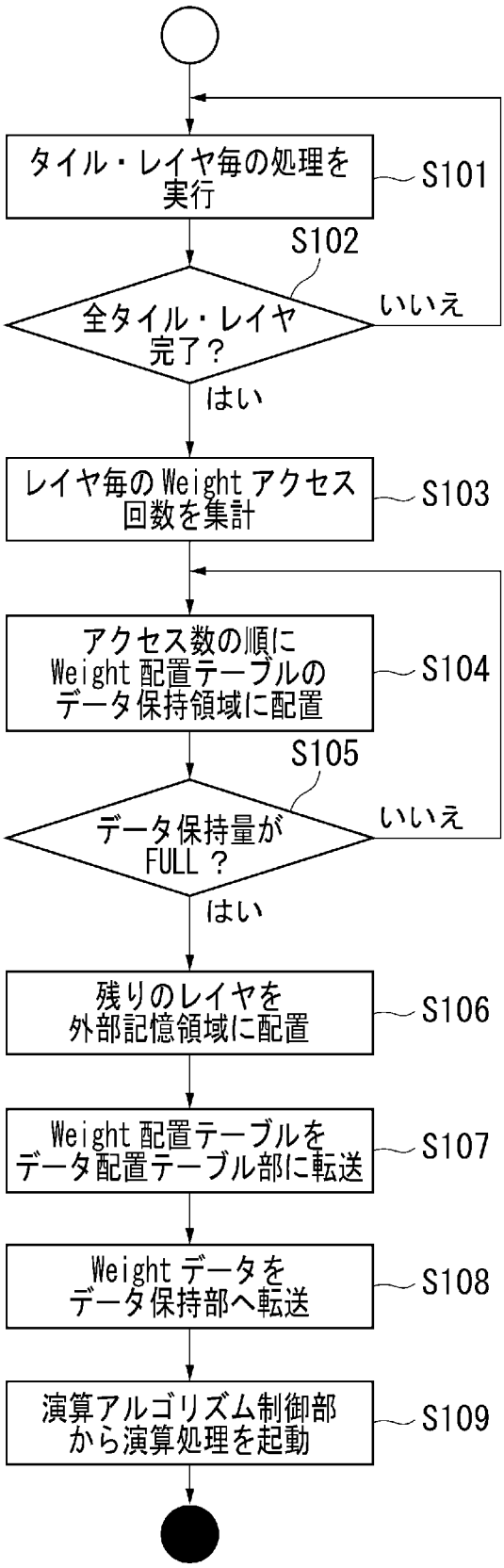
[図4]


5
507
508
506
読出し予測テーブル部
読出し予測テーブル制御部
読出しカウンタ部
セレクタ制御部
505
502
503
509
504
データ入力インターフェイス部
セレクタ部
データ出力インターフェイス部
データ伸長器
データ保持部
501
ローカルバッファ

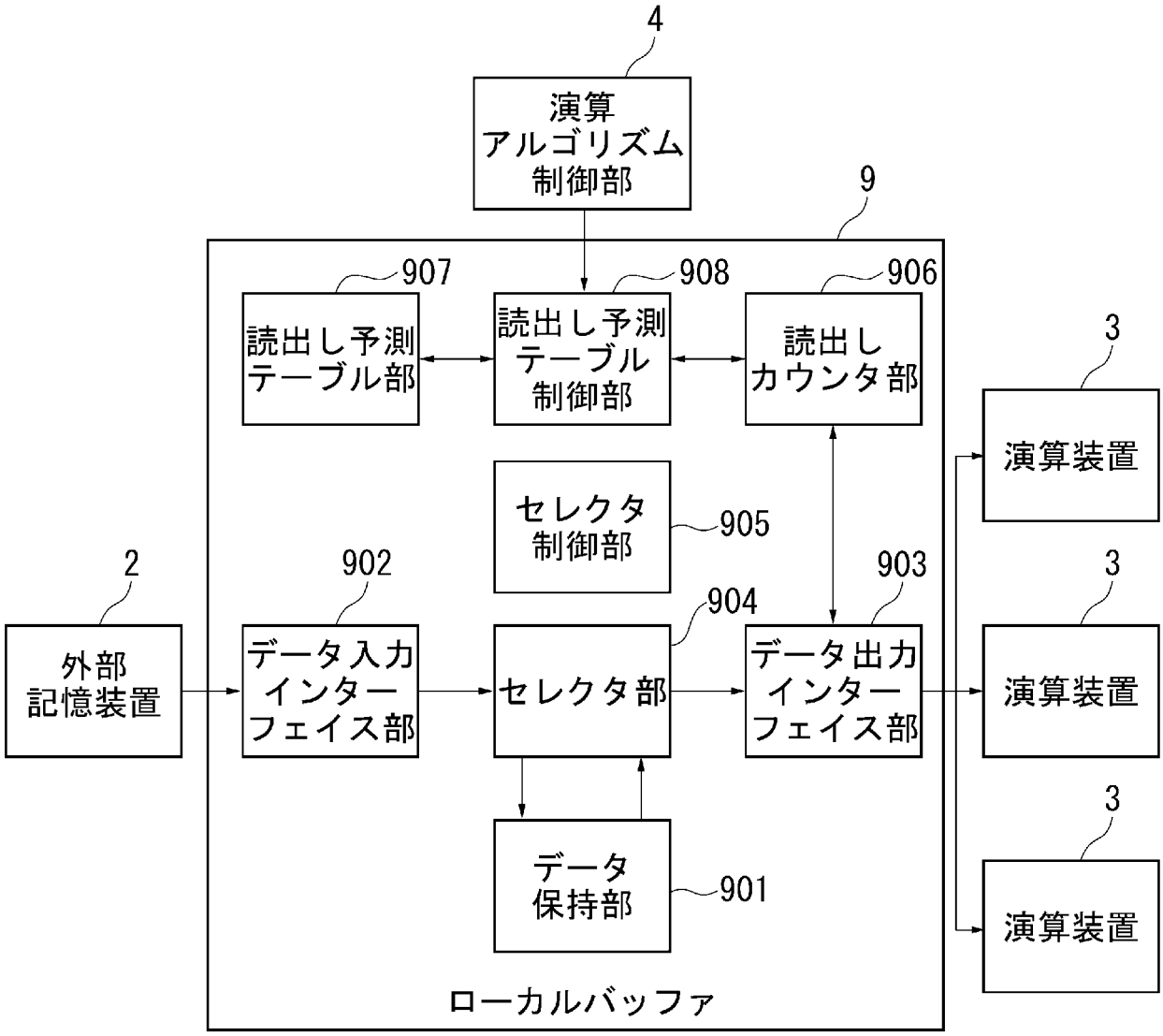
[図5]


4
演算
アルゴリズム
制御部
3
演算装置
6
603
データ出力
インター
フェイス部
606
データ配置
テーブル部
605
セレクタ
制御部
604
セレクタ部
601
データ
保持部
602
データ入力
インター
フェイス部
ローカルバッファ
2
外部
記憶装置
7
701
Weight 配置
テーブル作成手段
702
読出しカウント手段
Weight 配置プログラム
演算シミュレータ
8

[図6]


タイル・レイヤ毎の処理を実行
S101
S102
全タイル・レイヤ完了？
いいえ
はい
レイヤ毎のWeightアクセス回数を集計
S103
アクセス数の順にWeight配置テーブルのデータ保持領域に配置
S104
S105
データ保持量がFULL？
いいえ
はい
残りのレイヤを外部記憶領域に配置
S106
Weight配置テーブルをデータ配置テーブル部に転送
S107
Weightデータをデータ保持部へ転送
S108
演算アルゴリズム制御部から演算処理を起動
S109

[図7]


4
演算
アルゴリズム
制御部
9
907
908
906
読出し予測
テーブル部
読出し予測
テーブル
制御部
読出し
カウンタ部
3
演算装置
セレクタ
制御部
905
2
902
903
904
外部
記憶装置
データ入力
インター
フェイス部
セレクタ部
データ出力
インター
フェイス部
3
演算装置
3
データ
保持部
901
演算装置
ローカルバッファ

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2022/029689

A. CLASSIFICATION OF SUBJECT MATTER

G06F 12/126(2016.01)i; ***G06F 12/0866***(2016.01)i; ***G06F 12/0875***(2016.01)i; ***G06F 13/14***(2006.01)i
FI: G06F12/126 105; G06F12/0866 100; G06F12/0875 100; G06F13/14 310F

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)

G06F12/126; G06F12/0866; G06F12/0875; G06F13/14

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 1971-2022
Registered utility model specifications of Japan 1996-2022
Published registered utility model applications of Japan 1994-2022

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.
X	JP 2004-192403 A (FUJI XEROX CO LTD) 08 July 2004 (2004-07-08) abstract, paragraphs [0022]-[0027], [0033]-[0038], fig. 1, 3, 5	11
Y	abstract, paragraphs [0022]-[0027], [0033]-[0038], fig. 1, 3, 5	1-2, 4, 6-7, 9
A	entire text, all drawings	3, 5, 8, 10, 12
Y	JP 8-44625 A (NEC SOFTWARE LTD) 16 February 1996 (1996-02-16) paragraphs [0010]-[0011], fig. 1	1-2, 4, 6-7, 9
A	JP 2019-114013 A (FUJITSU ADVANCED ENG LTD) 11 July 2019 (2019-07-11) entire text, all drawings	1-12
A	JP 2015-525940 A (INTEL CORPORATION) 07 September 2015 (2015-09-07) entire text, all drawings	1-12
A	US 2020/0327061 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 15 October 2020 (2020-10-15) entire text, all drawings	1-12

☑ 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
16 September 2022	**27 September 2022**
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

International application No.
PCT/JP2022/029689

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2020-140507 A (NEC PLATFORMS LTD) 03 September 2020 (2020-09-03) entire text, all drawings	1-12

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2022/029689

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP 2004-192403 A	08 July 2004	(Family: none)	
JP 8-44625 A	16 February 1996	(Family: none)	
JP 2019-114013 A	11 July 2019	(Family: none)	
JP 2015-525940 A	07 September 2015	US 2014/0095778 A1 entire text, all drawings WO 2014/052157 A1 KR 10-2015-0036176 A CN 104662519 A	
US 2020/0327061 A1	15 October 2020	WO 2019/127487 A1 entire text, all drawings CN 110226158 A	
JP 2020-140507 A	03 September 2020	(Family: none)	

国際調査報告

国際出願番号
PCT/JP2022/029689

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06F 12/126(2016.01)i; G06F 12/0866(2016.01)i; G06F 12/0875(2016.01)i; G06F 13/14(2006.01)i
FI: G06F12/126 105; G06F12/0866 100; G06F12/0875 100; G06F13/14 310F

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））

G06F12/126; G06F12/0866; G06F12/0875; G06F13/14

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	１９７１－２０２２年
日本国実用新案登録公報	１９９６－２０２２年
日本国登録実用新案公報	１９９４－２０２２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
X	JP 2004-192403 A (富士ゼロックス株式会社) 08.07.2004 (2004-07-08) [要約], 段落[0022]-[0027],[0033]-[0038], 図1,3,5	11
Y	[要約], 段落[0022]-[0027],[0033]-[0038], 図1,3,5	1-2,4,6-7,9
A	全文, 全図	3,5,8,10,12
Y	JP 8-44625 A (日本電気ソフトウェア株式会社) 16.02.1996 (1996-02-16) 段落[0010]-[0011], 図1	1-2,4,6-7,9
A	JP 2019-114013 A (株式会社富士通アドバンストエンジニアリング) 11.07.2019 (2019-07-11) 全文, 全図	1-12
A	JP 2015-525940 A (インテル コーポレイション) 07.09.2015 (2015-09-07) 全文, 全図	1-12
A	US 2020/0327061 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 15.10.2020 (2020-10-15) 全文, 全図	1-12

☑ 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
16.09.2022	27.09.2022

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）
北村　学 5S 4535

電話番号 03-3581-1101 内線 3546

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
A	JP 2020-140507 A（ＮＥＣプラットフォームズ株式会社）03.09.2020（2020-09-03） 全文，全図	1-12

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/JP2022/029689

引用文献	公表日	パテントファミリー文献	公表日
JP 2004-192403 A	08.07.2004	（ファミリーなし）	
JP 8-44625 A	16.02.1996	（ファミリーなし）	
JP 2019-114013 A	11.07.2019	（ファミリーなし）	
JP 2015-525940 A	07.09.2015	US 2014/0095778 A1 全文，全図 WO 2014/052157 A1 KR 10-2015-0036176 A CN 104662519 A	
US 2020/0327061 A1	15.10.2020	WO 2019/127487 A1 全文，全図 CN 110226158 A	
JP 2020-140507 A	03.09.2020	（ファミリーなし）	