(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2021-108104
(P2021-108104A)
(43)公開日　令和3年7月29日(2021.7.29)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 15/80	(2006.01)	G06F	15/80		5B056
G06N 3/063	(2006.01)	G06N	3/063		
G06F 17/16	(2006.01)	G06F	17/16	M	

審査請求 未請求　請求項の数 20　ＯＬ　外国語出願　（全 25 頁）

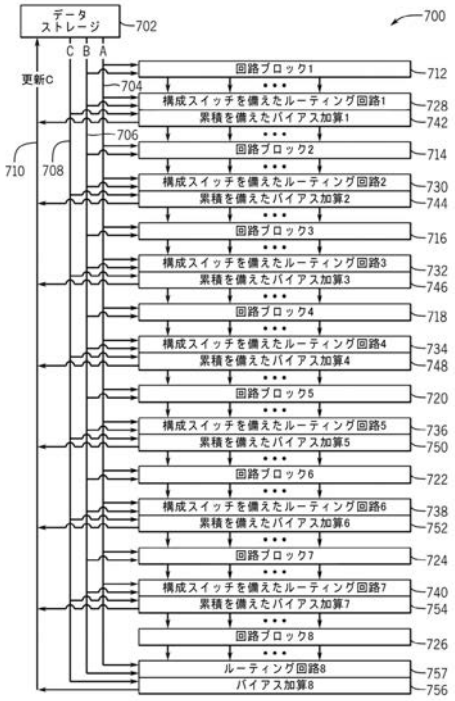
(21)出願番号　特願2020-157941 (P2020-157941)
(22)出願日　令和2年9月18日 (2020.9.18)
(31)優先権主張番号　16/729,381
(32)優先日　令和1年12月28日 (2019.12.28)
(33)優先権主張国・地域又は機関
米国 (US)

(71)出願人　591003943
インテル・コーポレーション
アメリカ合衆国　９５０５４　カリフォルニア州・サンタクララ・ミッション　カレッジ　ブーレバード・２２００
(74)代理人　110000877
龍華国際特許業務法人
(72)発明者　カムレシュ　アール．ピレイ
アメリカ合衆国　９５０５４　カリフォルニア州・サンタクララ・ミッション　カレッジ　ブーレバード・２２００　インテル・コーポレーション内

最終頁に続く

(54)【発明の名称】部分的読み取り／書き込みが可能な再構成可能なシストリックアレイのシステム及び方法

(57)【要約】　（修正有）
【課題】再構成可能なシストリックアレイ回路を含むデータ処理システムを提供する。
【解決手段】再構成可能なシストリックアレイ回路７００は、１又は複数のグループの処理要素を含む第１回路ブロック７１２と、１又は複数のグループの処理要素を含む第２回路ブロック７１４とを含む。再構成可能なシストリックアレイ回路は、さらに、累積値、乗算積、又はそれらの組み合わせに行列バイアスを加算する累積回路を備えた第１バイアス加算７４２を含む。再構成可能なシストリックアレイ回路は、さらに、第１回路ブロックから第２回路ブロックへ、第１回路ブロックから累積回路を備えた第１バイアス加算へ又はそれらの組み合わせへ微分をルーティングするた第１ルーティング回路７２８をさらに含む。
【選択図】図７



【特許請求の範囲】
【請求項１】
データを格納するように構成されたデータストレージと、
再構成可能なシストリックアレイ回路と、を備え、前記再構成可能なシストリックアレイ回路は、
前記データを処理するように構成された１又は複数のグループの処理要素を有する第１回路ブロックと、
前記データを処理するように構成された１又は複数のグループの処理要素を有する第２回路ブロックと、
行列バイアスを累積値に、乗算積に、又はそれらの組み合わせに加算するように構成された累積回路を備えた第１バイアス加算と、
微分を前記第１回路ブロックから前記第２回路ブロックに、前記第１回路ブロックから累積回路を備えた前記第１バイアス加算に、又はそれらの組み合わせにルーティングするように構成された第１ルーティング回路と、
を含む、システム。
【請求項２】
前記第１ルーティング回路は、構成スイッチ信号を受信することに基づいて、前記微分を前記第１回路ブロックから前記第２回路ブロックに、前記第１回路ブロックから累積回路を備えた前記第１バイアス加算に、又はそれらの組み合わせにルーティングするように構成されたデマルチプレクサ及びマルチプレクサ回路を含む、請求項１に記載のシステム。
【請求項３】
累積回路を備えた前記第１バイアス加算は、クロック信号に基づいて前記乗算積を前記累積値として累積するように構成された記憶回路、及び前記行列バイアスを前記累積値に、前記乗算積に、又はそれらの組み合わせに加算するように構成された少なくとも１つの加算器を含む、請求項１又は２に記載のシステム。
【請求項４】
累積回路を備えた前記第１バイアス加算は、Ｎの加算器のレイテンシを含み、前記記憶回路は、Ｎ個の記憶コンポーネントを含む、請求項３に記載のシステム。
【請求項５】
前記Ｎ個の記憶コンポーネントは、それぞれ、フリップフロップを含む、請求項４に記載のシステム。
【請求項６】
前記記憶回路は、前記Ｎ個の記憶コンポーネントをマルチプレクサに結合するＮ個のラインを含み、前記記憶回路は、前記加算器のレイテンシが演算中にＮを超えた場合に、累積値を前記Ｎ個の記憶コンポーネントから前記マルチプレクサに前記Ｎ個のラインを介して送信するように構成される、請求項４又は５に記載のシステム。
【請求項７】
累積回路を備えた前記第１バイアス加算は、累積回路を備えた前記第１バイアス加算に入る新しい値を前記累積値に加算し、前記加算の結果を前記Ｎ個の記憶コンポーネントに格納するように構成される、請求項６に記載のシステム。
【請求項８】
１又は複数のグループの処理要素を有する第３回路ブロックと、
第２行列バイアスを第２累積値に、前記乗算積に、又はそれらの組み合わせに加算するように構成された累積回路を備えた第２バイアス加算と、
微分を前記第２回路ブロックから前記第３回路ブロックに、前記第２回路ブロックから累積回路を備えた前記第２バイアス加算に、又はそれらの組み合わせにルーティングするように構成された第２ルーティング回路と、
を備える、請求項１から７のいずれか一項に記載のシステム。
【請求項９】

前記第３回路ブロックの下流に配置され、第３行列バイアスを前記第３回路ブロックから出力に加算するように構成されたバイアス加算回路を備える、請求項８に記載のシステム。

【請求項１０】

前記再構成可能なシストリックアレイ回路を使用するよう、又は前記再構成可能なシストリックアレイ回路を含むように構成されたホストプロセッサ（ＣＰＵ）を備え、

前記ＣＰＵは、『「Ｍ」累積によるタイル部分「Ｎ」ドット積』命令であり、前記Ｎは、ともにマージされている異なる行列の数であり、Ｍは、前記再構成可能なシストリックアレイ回路への入力として使用される不完全な行列の数である、前記命令、前記再構成可能なシストリックアレイ回路への入力として使用されるともにマージされている前記異なる行列のサイズを指定する即値を有する「ドット積に対するタイルサイズ」命令、累積回路を備えた前記第１バイアス加算を制御する「タイル累積ドット積」命令、又はそれらの組み合わせを実行するように構成される、

請求項１から９のいずれか一項に記載のシステム。

【請求項１１】

行列Ａ及び行列Ｂに基づいてデータの１又は複数のタイルのそれぞれに対してタイルサイズを決定する段階と、

タイルサイズに基づいて、完全タイル、不完全タイル、又はそれらの組み合わせを導出する段階と、

行列Ｃの結果を導出するために、前記完全タイル、前記不完全タイル、又はそれらの組み合わせを、再構成可能なシストリックアレイ回路を介して処理する段階であり、前記完全タイル、前記不完全タイル、又はそれらの組み合わせを処理する段階とは、前記行列Ｃの結果を提供するために、前記再構成可能なシストリックアレイ回路に含まれるルーティング回路及び前記再構成可能なシストリックアレイ回路に含まれる累積回路を備えたバイアス加算又はそれらの組み合わせを適用することを含む、段階と、

を備える方法。

【請求項１２】

前記再構成可能なシストリックアレイ回路は、Ｎ行Ｍ列のアレイサイズを含み、前記完全タイルは、Ｎ行以下及びＭ列以下を有する完全サイズを含み、前記不完全タイルは、Ｎ行より多く、Ｍ列より多く、又はそれらの組み合わせを有する不完全サイズを含む、請求項１１に記載の方法。

【請求項１３】

前記ルーティング回路を適用する段階は、微分を１又は複数のグループの処理要素を含む第１回路ブロックから１又は複数のグループの処理要素を含む第２回路ブロックにルーティングする段階、及び微分を前記第１回路ブロックから累積回路を備えた前記バイアス加算に、又はそれらの組み合わせにルーティングする段階を含み、

微分を前記第１回路ブロックから累積回路を備えた前記バイアス加算にルーティングする段階は、前記微分を累積回路を備えた前記バイアス加算にて受信し、前記微分を行列Ｃのバイアスに加算するために累積値に累積する段階を含む、請求項１１又は１２に記載の方法。

【請求項１４】

前記完全タイル、前記不完全タイル、又はそれらの組み合わせを前記再構成可能なシストリックアレイ回路を介して処理する段階は、

行列Ｃのアドレス衝突を検出し、累積回路を備えた前記バイアス加算に通信される累積イネーブル信号を自動的にオンにするように構成されたマイクロアーキテクチャモードを適用する段階と、

前記再構成可能なシストリックアレイ回路への入力として使用されるともにマージされている異なる行列のサイズを指定する即値を有する「ドット積に対するタイルサイズ」命令、累積回路を備えた前記バイアス加算を制御する「タイル累積ドット積」命令、又はそれらの組み合わせを実行することによりアーキテクチャモードを適用する段階と、

を含む、請求項１１から１３のいずれか一項に記載の方法。
【請求項１５】
データを格納するように構成されたデータストレージと、
再構成可能なシストリックアレイ回路と、
前記再構成可能なシストリックアレイ回路に接続されるコアのデコーダであり、単一の命令をデコードされた１又は複数の命令にデコードする、デコーダと、を備え、
前記１又は複数の命令は、
行列Ａ及び行列Ｂを表す前記データを、前記データストレージから、前記データを処理し、前記データに基づく微分を提供するように構成された１又は複数のグループの処理要素を含む第１回路ブロックに通信し、
再構成可能なルーティング回路のスイッチングオン又はオフに基づいて、前記微分を前記第１回路ブロックから第２回路ブロックに、累積回路を備えたバイアス加算に、又はそれらの組み合わせにルーティングするように構成され、
累積回路を備えた前記バイアス加算は、行列バイアスを、累積値に、行列Ａ及び行列Ｂの乗算積に、又はそれらの組み合わせに加算するように構成され、
前記第１回路ブロック、前記第２回路ブロック、前記再構成可能なルーティング回路、累積回路を備えた前記バイアス加算、又はそれらの組み合わせは、前記再構成可能なシストリックアレイ回路に含まれる、装置。
【請求項１６】
前記単一の命令は、デコードされると、前記再構成可能なシストリックアレイ回路への入力として使用されるとともにマージされている異なる行列のサイズを指定する即値を有する「ドット積に対するタイルサイズ」命令、累積回路を備えた前記バイアス加算を制御する「タイル累積ドット積」命令、又はそれらの組み合わせを介してアーキテクチャモードを使用する、請求項１５に記載の装置。
【請求項１７】
前記単一の命令は、『「Ｍ」累積によるタイル部分「Ｎ」ドット積』命令を含み、Ｎは、ともにマージされている異なる行列の数であり、Ｍは、前記再構成可能なシストリックアレイ回路への入力として使用される不完全な行列の数である、請求項１６に記載の装置。
【請求項１８】
前記単一の命令は、デコードされると、前記再構成可能なシストリックアレイ回路に、前記データを使用することによってＣ＝＋Ａ＊Ｂを解かせ、前記データは、前記行列Ａ及び前記行列Ｂを表す、請求項１５から１７のいずれか一項に記載の装置。
【請求項１９】
前記再構成可能なシストリックアレイ回路を有する回路を備え、
前記回路は、マイクロプロセッサ、ハードウェアアクセラレータ、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）、特定用途向け集積回路（ＡＳＩＣ）、カスタムマイクロチップ、又はそれらの組み合わせを含む、請求項１５から１８のいずれか一項に記載の装置。
【請求項２０】
行列Ａ及び行列Ｂに基づいてデータの１又は複数のタイルのそれぞれに対してタイルサイズを決定する手段と、
タイルサイズに基づいて、完全タイル、不完全タイル、又はそれらの組み合わせを導出する手段と、
行列Ｃの結果を導出するために、前記完全タイル、前記不完全タイル、又はそれらの組み合わせを、再構成可能なシストリックアレイ回路を介して処理する手段であり、前記完全タイル、前記不完全タイル、又はそれらの組み合わせを処理することは、前記行列Ｃの結果を提供するために、前記再構成可能なシストリックアレイ回路に含まれるルーティング回路及び前記再構成可能なシストリックアレイ回路に含まれる累積回路を備えたバイアス加算又はそれらの組み合わせを適用することを含む、手段と、

を備えるシステム。
【発明の詳細な説明】
【背景技術】
【０００１】
本開示は、概して、シストリックアレイベースのアクセラレータに関し、より詳細には、部分的読み取り／書き込みを伴うシストリックアレイベースのアクセラレータに関する。
【０００２】
本セクションは、読者に本開示の様々な態様に関連し得る技術分野の様々な態様を紹介することが意図されており、当該態様は、以下に説明される及び／又は特許請求の範囲に記載される。この検討は、本開示の様々な態様のより良好な理解を容易にするための背景情報を読者に提供することに役立つと考えられる。従って、これらの記述は、先行技術を承認するものとしてではなく、この観点で読むべきものであることを理解されたい。
【０００３】
シストリックアレイベースのアクセラレータの使用は、ディープニューラルネットワーク（ＤＮＮ）ベースのアプリケーションで有用なそれらのようなより効率的な計算を提供してよい。シストリックアレイベースのＤＮＮアクセラレータは、アプリケーションの計算エンジンを提供するために、数百の算術演算ユニット、例えば処理要素（ＰＥ）を使用してよい。ＤＮＮアクセラレータは、通常且つ固定サイズの密行列乗算に対してより最適化されてよい。例えば、算術演算ユニットのシストリックアレイ実装が使用されて、性能を向上し、表面積を減らし、電力の利益を得てよい。従って、特定のＤＮＮアクセラレータは、非常に規則的なデータフロー用に最適化された高密度２次元（２Ｄ）アレイを使用してよい。多くのＤＮＮアクセラレータは、比較的低速又は非効率であり得る。
【図面の簡単な説明】
【０００４】
【図１】本開示の実施形態による、再構成可能なシストリックアレイベースのアクセラレータの回路を有する１又は複数のプロセッサを含むデータ処理システムのブロック図である。
【０００５】
【図２】本開示の実施形態によるシストリックアレイシステムの例のブロック図である。
【０００６】
【図３】本開示の実施形態による部分的なバイアス累積サポートを含む再構成可能なシストリックアレイシステムを実行するために使用され得るスケジューラの実施形態のブロック図である。
【０００７】
【図４】本開示の実施形態による、図３の再構成可能なシストリックアレイシステムのさらに詳細を示すブロック図である。
【０００８】
【図５】本開示の実施形態による、再構成可能なルーティング回路及び累積回路を備えたバイアス加算の実施形態を示す概略図である。
【０００９】
【図６】本開示の実施形態による、さらに詳細を示す累積回路を備えたバイアス加算の実施形態の概略図である。
【００１０】
【図７】本開示の実施形態による、複数の再構成可能なルーティング回路及び累積回路を備えたバイアス加算を有する再構成可能なシストリックアレイシステムを示すブロック図である。
【００１１】
【図８】本開示の実施形態による再構成可能なシストリックアレイシステムの回路を実行するのに好適な処理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】
【００１２】
１又は複数の具体的な実施形態が以下に説明される。これらの実施形態の簡潔な記載を提供するために、実際の実装のすべての特徴が本明細書において説明されるわけではない。任意のエンジニアリング又は設計プロジェクトにおけるように、任意のそのような実際の実装の開発においては、システム関連及びビジネス関連の制約の順守などの開発者の特定の目的を達成すべく、多数の実装固有の決定がなされなければならず、当該制約は実装によって変動され得ることを理解されたい。さらに、そのような開発への取り組みは複雑且つ時間を消費するものであり得るが、それにもかかわらず、本開示の利益を有する当業者には設計、製造、及び製造のルーチン作業であることを理解されたい。
【００１３】
本明細書で説明される技術は、ディープニューラルネットワーク（ＤＮＮ）で使用されるそれらの計算のような特定の計算を向上するのに有用な特定のシストリックアレイ技術を含む。シストリックアレイは、処理ユニットがセル又はノードと称されてよい密に接続された処理ユニットの同種のネットワークを含んでよい。各ノードには、様々な計算を提供するために使用されてよい融合型積和ユニット（ＦＭＡ）のような処理要素（ＰＥ）を含んでよい。データはシストリックアレイに入り、アレイのＦＭＡを通じて、例えば隣接するＦＭＡ間を流れてよく、データフローの結果は、特定のアプリケーション、例えばＤＮＮアプリケーションの計算として提供されてよい。ＤＮＮシストリックアレイアクセラレータは、通常且つ固定サイズの密行列乗算に対してより最適化されてよい。例えば、ＤＮＮシストリックアレイアクセラレータは、非常に規則的なデータフロー用により最適化された高密度２次元アレイを使用してよい。ＤＮＮシストリックアレイアクセラレータを介して解決される非常に大きい又は小さいのいずれかの問題及び／又は提供された規則的データフローにうまくマッピングされない問題は、部分的な結果の複数の読み取り／書き込み、及び／又はシストリックアレイ内のＰＥの大きく低下した使用率を引き起こし得る。
【００１４】
ディープラーニングアプリケーションは、密ＤＮＮと疎ＤＮＮを含むように分類されてよい。密及び疎ＤＮＮの両方に対して、実行のいくつかの部分は、所与のシストリックアレイに対する規則的なデータフロー上に完全にマッピングされてよいが、すべてではない。例えば、密ＤＮＮの場合、問題のサイズは非常に大きくなってよく、行列Ａ、Ｂ、及びＣの配列計算がＣ＋＝Ａ＊Ｂのような方程式を含む場合、各行列は複数のタイル（例えば、２Ｄデータ構造）に分割されて、行列をシストリックアレイに「合致」させてよい。例えば、Ｘ次元にｘＰＥ及びＹ次元にｙＰＥを有するシストリックアレイにおいてＣの単一のタイルを計算するために、Ａに対するＸ次元及びＢに対するＹ次元におけるすべての対応するｘ、ｙタイルに沿った計算が使用されてよい。Ｘ及びＹ次元の計算は、各それぞれのタイル乗算から生成される部分的な結果が書き出され、次いで、累積の単一の「チェーン」内のすべてのタイルの完了まで、さらなる処理（例えば、他の部分的な結果との累積）のために読み戻すことを必要としてよい。行列はタイルより小さくてよく（例えば、すべてのタイルよりも少ないスペースを使用する）、行列はタイルと同一のサイズであってよく、又は行列は複数のタイルを使用してよい（例えば、行列はいずれか１つのタイルのサイズより大きい）ことに留意されたい。従って、本明細書で称されるタイル又はタイルデータは、Ｎ列Ｍ行、いくつかの場合にＮ＝Ｍを有するデータのアレイを含んでよい。行及び／又は列は、本明細書で「グループ」と称されてよい。
【００１５】
疎ＤＮＮの場合、「ブロックスパーシティ」処理が存在してよく、この場合に行列は任意のサイズの密なブロックにより表される。そのような密なブロック表現は、ゼロが表される必要がなくてよいので、行列内の多くの又はほとんどのゼロに亘って「スキップ」することを可能にしてよい。しかし、ブロックスパーシティの副作用は、一般的な行列乗算（ＧＥＭＭ）の微分のような特定の微分を計算する場合に、入力行列に小さい及び／又は

不規則なサイズのブロックが見つけられる可能性がある。すべてのディープラーニングアプリケーション（疎並びに密）に対して、部分的読み取り／書き込みを縮小し、シストリックアレイ上で不規則な幅で複数の行列乗算を実行し、それによりＰＥの利用が高くなることは有益であろう。さらに、現在、シストリックアレイのサイズに完全に合致するであろう密行列乗算に対する性能、面積、及び／又は電力不利益の支払いを最小に抑えつつ、ＰＥの利用を向上することは有益であろう。

【００１６】

本明細書で説明される技術は、部分的な累積サポートを備えた再構成可能なシストリックアレイを含む。累積サポートは、既存のタイルストレージから分離した、例えばスケジューラを介して複数の行列乗算を処理するのに好適な累積ストレージを含んでよい。スケジューラは、行列が実行のために送信される順序をスケジューリングしてよく、新しい命令（複数可）（例えば、マクロ命令）が使用されて、再構成可能なシストリックアレイを通じてデータフローを実行してよい。２つ（又はより多く）の命令が、さらに下で説明されるように計算間において宛先を使用（又は上書き）しないで同一の宛先を有する場合に、複数の行列乗算命令に亘ってシストリックアレイの宛先をチェックし、ソフトウェアの介入なしにハードウェアベースの計算を有効にするために使用されるマイクロアーキテクチャ機能が提供されてよい。再構成可能なシストリックアレイは、スケジューラによってスケジューリングされているタイルに基づいて有効にされてよい累積ロジックシステムを含む。累積ロジックシステムは、解かれる問題の終わりまで部分値を累積し、最終出力をストレージ（例えば、メモリ、バッファ、レジスタなど）に書き込んでよい。再構成可能なシストリックアレイを提供することにより、ハードウェアベースの計算がより柔軟になってよく、それとともに、ハードウェア及びストレージ（例えば、タイルレジスタファイル）間のデータ転送がさらに削減されるため、ＤＮＮベースのアプリケーションのような特定のアプリケーションに対する利用を向上し、データ転送を減少してよい。

【００１７】

上記を念頭に置いて、図１は、本開示の実施形態による、１又は複数のプロセッサ（複数可）１０２を含むデータ処理システム１００のブロック図である。データ処理システム１００は、示されているものより多い又はより少ないコンポーネント（例えば、電子ディスプレイ、ユーザインタフェース構造、特定用途向け集積回路（ＡＳＩＣ））を含んでよい。データ処理システム１００は、データ処理システム１００へのデータ処理要求（例えば、ＤＮＮ計算、機械学習、ビデオ処理、音声認識、画像認識、データ圧縮、データベース検索ランキング、バイオインフォマティクス、ネットワークセキュリティパターンの認識、空間ナビゲーションなどを実行する）を管理してよいＩＮＴＥＬ（登録商標）第１０世代プロセッサ（例えば、アイスレークプロセッサ）のような１又は複数のプロセッサ１０２を介して特定のコード又はコンピュータ命令を実行してよい。本明細書における命令という用語は、マクロ命令、例えば実行のためにプロセッサ１０２に提供される命令、又はマイクロ命令、例えばマクロ命令をデコードするプロセッサ１０２のデコーダから得られる命令を指してよいことに留意されたい。デコーダは、プロセッサ１０２のコアに含まれてよい。

【００１８】

プロセッサ（複数可）１０２は、メモリ及び／又は記憶回路１０４と通信し得る。メモリ及び／又は記憶回路１０４は、ランダムアクセスメモリ（ＲＡＭ）、リードオンリメモリ（ＲＯＭ）、１又は複数のハードドライブ、フラッシュメモリ、又は任意の他の好適な光、磁気、若しくはソリッドステート記憶媒体のような有形で非一時的な機械可読媒体であってよい。メモリ及び／又は記憶回路１０４は、プロセッサ実行可能制御ソフトウェア、構成ソフトウェア、システムパラメータ、構成データ等のようなデータ処理システム１００により処理されるデータを保持してよい。

【００１９】

また、データ処理システム１００は、データ処理システム１００が他の電子デバイスと通信することを許可するネットワークインタフェース１０６を含んでよい。いくつかの実

施形態では、データ処理システム１００は、様々な異なる要求を処理するデータセンタの一部であってよい。例えば、データ処理システム１００は、ネットワークインタフェース１０６を介してデータ処理要求を受信して、ＤＮＮ計算、機械学習、ビデオ処理、音声認識、画像認識、データ圧縮、データベース検索ランキング、バイオインフォマティクス、ネットワークセキュリティパターンの認識、空間ナビゲーション、又はいくつかの他の専門なタスクを実行してよい。また、データ処理システム１００は、ディスプレイデバイス（例えば、コンピュータモニタ）、キーボード、マウス、スピーカ、音声入力デバイスなどのような、情報の入力及び／又は表示に有用な１又は複数の入力／出力システム１０８を含んでよい。

【００２０】

示される実施形態では、プロセッサ１０２は、動作可能及び／又は通信可能に再構成可能なシストリックアレイシステム１１０に接続されてよい。再構成可能なシストリックアレイシステム１１０は、複数の処理要素（ＰＥ）、及び再構成可能なシストリックアレイシステム１１０のＰＥのいくつか（又はすべて）を通じてデータ（例えば、データフロー）を再構成可能に移動するために使用されてよい再構成可能なルーティングシステム１１２を含む、データをルーティングするのに好適な特定の回路を含んでよい。従って、ＤＮＮアプリケーションに使用されるデータのようなデータは、例えば、プロセッサ１０２を介して再構成可能なシストリックアレイシステム１１０に提供されてよく、次いで、再構成可能なシストリックアレイシステム１１０は、例えば、再構成可能なルーティングシステム１１２を介して、以下でさらに説明するように改善されたデータフローをより柔軟に導出してよい。再構成可能なシストリックアレイシステム１１０は、さらに、特定のバイアスデータを累積及び加算するのに好適な、累積システム１１４を備えるバイアス加算を含んでよい。例えば、累積システム１１４を備えたバイアス加算は、解かれる問題の終わりまで部分的な計算値（例えば、行列バイアス値）を累積し、最終出力をストレージに書き込んでよい。

【００２１】

シストリックアレイシステムを記載することは有益であってよい。ここで図２を参照すると、図は、シストリックアレイシステム２００の処理要素（ＰＥ）を通じたデータフローを介して、ＤＮＮベースの問題のような特定の問題を解くために使用されてよいシストリックアレイシステム又は回路２００を示すブロック図である。例えば、シストリックアレイシステム２００は、Ｃ＋＝Ａ＊Ｂ（例えば、更新

【数１】

$$c_{i,j} = c_{i,j} + \sum_{l=0}^{K-1} a_{i,l} * b_{l,j}$$

ただし、Ｋは行列の行高さである）のような様々な計算を計算するのに使用されてよい。

【００２２】

示される実施形態では、データストレージ（例えば、複数のレジスタ、キャッシュ、バッファ等を有するレジスタファイル）２０２は、タイルデータのような行列Ａ、Ｂ、Ｃのデータを格納するのに使用されてよい。データストレージは、ライン２０４、２０６、２０８、及び２１０を使用して、行列Ａのタイルデータ、行列Ｂのタイルデータ、行列Ｃのタイルデータ、更新された行列Ｃのタイルデータをそれぞれ通信してよい。ライン２０４、２０６、２０８、及び２１０のそれぞれは、複数の導管を含んでよいことに留意されたい。すなわち、ライン２０４、２０６、２０８、及び２１０は、それぞれポートであってよく、各ポートは複数の導管又はラインを有してよい。ルーティング回路２１２は、行列Ａの行０及び列０に対応する値Ａ［０］［０］を受信してよく、次いで、ルーティング回路２１２は、第１値Ａ［０］［０］を、処理要素２１４、２１６、２１８などのようなシストリックアレイシステム２００の第１行内の処理要素にブロードキャストしてよい。ルーティング回路２１２は、さらに、Ｂ内の第１行の値を表す値Ｂ［０］［０］、Ｂ［０］［１］、Ｂ［０］［２］、…、Ｂ［０］［Ｋ］を受信して、値を処理要素２１４、２１６

、２１８などにブロードキャストしてよい。例えば、処理要素２１４は、値Ｂ［０］［０］を受信してよく、処理要素２１６は、値Ｂ［０］［１］を受信してよく、処理要素２１８は、値Ｂ［０］［Ｋ］を受信してよい。次いで、所与の行に対する処理要素のいくつか又はすべては、受信した入力に基づいて、乗算及び加算演算のような特定の演算の結果を出力してよい。次いで、例えば、処理要素２１４は、乗算Ａ［０］［０］＊Ｂ［０］［０］の結果を出力してよく、処理要素２１６は、乗算Ａ［０］［０］＊Ｂ［０］［１］の結果を出力してよく、処理要素２１８は、乗算Ａ［０］［０］＊Ｂ［０］［Ｋ］の結果を出力してよい。次いで、処理要素２１４、２１６、２１８の出力は、ルーティング回路２２０に送信されてよい。

【００２３】

ルーティング回路２２０は、行列Ａの行０及び列１に対応する値Ａ［０］［１］を受信してよく、次いで、ルーティング回路２２０は、値Ａ［０］［１］を、処理要素２２２、２２４、２２６などのようなシストリックアレイシステム２００の第２行内の処理要素にブロードキャストしてよい。同様に、ルーティング回路２２０は、さらに、Ｂ内の第２行の値を表す値Ｂ［１］［０］、Ｂ［１］［１］、Ｂ［１］［２］、…、Ｂ［１］［Ｋ］を受信して、値を処理要素２２２、２２４、２２６などにブロードキャストしてよい。例えば、処理要素２２２は、値Ｂ［１］［０］を受信してよく、処理要素２２４は、値Ｂ［１］［１］を受信してよく、処理要素２２６は、値Ｂ［１］［Ｋ］を受信してよい。次いで、所与の行に対する処理要素のいくつか又はすべては、受信した入力に基づいて、乗算演算のような特定の演算の結果を出力してよい。次いで、例えば、処理要素２２２は、処理要素２１４の出力に加算された乗算Ａ［０］［１］＊Ｂ［１］［０］の結果を出力して、出力Ａ［０］［１］＊Ｂ［１］［０］＋Ａ［０］［０］＊Ｂ［０］［０］に到達してよい。次いで、同様に、処理要素２２４は、処理要素２１６の出力に加算された乗算Ａ［０］［１］＊Ｂ［１］［１］の結果を出力して、出力Ａ［０］［１］＊Ｂ［１］［１］＋Ａ［０］［０］＊Ｂ［０］［１］に到達してよい。次いで、同様に、処理要素２２６は、処理要素２１８の出力に加算された乗算Ａ［０］［１］＊Ｂ［１］［Ｋ］の結果を出力して、出力Ａ［０］［１］＊Ｂ［１］［Ｋ］＋Ａ［０］［０］＊Ｂ［０］［Ｋ］に到達してよい。そのような積和演算は、融合型積和と称されてよく、各処理要素に含まれる融合型積和ユニット（ＦＭＡ）を使用してよい。次いで、処理要素２２２、２２４、２２６の出力は、ルーティング回路２２８に送信されてよい。

【００２４】

同様の態様で、ルーティング回路２２８及び２３０は、行列Ａのデータ Ａ［０］［２］及びＡ［０］［３］をそれぞれ受信し、データをそれらのそれぞれの行の処理要素、例えば、ルーティング回路２２８に対する処理要素２３２、２３４、２３６及びルーティング回路２３０に対する処理要素２３８、２４０、２４２にブロードキャストしてよい。同様に、ルーティング回路２２８及び２３０は、行列の第３及び第４行に対する行列Ｂのデータを受信して、第３行のデータを処理要素２３２、２３４、２３６に、第４行のデータを処理要素２３８、２４０、２４２にそれぞれパスしてよい。また、処理要素２３２、２３４、２３６、２３８、２４０、及び２４２は、ＦＭＡ機能、従って、シストリックアレイシステム２００における行列Ａの入力、行列Ｂの入力、及び前の処理要素の出力を含む受信した入力に基づく上述のような乗算及び加算を提供してよい。実際、示されるすべての処理要素は、融合型積和ユニットを含んでよい。

【００２５】

次いで、バイアス加算回路２４４は、例えば、前に行列Ａ、Ｂに対して実行された演算、例えば、Ｃ＋＝Ａ＊Ｂ（例えば、行列Ｃからのバイアスを処理要素２３８、２４０、２４２からのそれぞれの結果に加算する）を用いて行列Ｃを加算及び／又は更新するために使用されてよい。例えば、ライン（複数可）２０８を介して受信された行列Ｃの値は、処理要素２３８、２４０、２４２などの出力に加算されて、ライン（複数可）２１０を介して更新行列Ｃとして格納されてよい。され得る。シストリックアレイシステム２００の実施形態は、４行の処理要素を有するものとして示されているが、他の実施形態は、より多

い行又はより少ない行を含んでよいことを理解されたい。特定の実施形態では、シストリックアレイシステム２００は、行ごとに３２個の処理要素を使用してよい。例えば、１２８の行列幅を有する密ＤＮＮワークロードを処理する場合、処理される行列は、タイルごとに３２列を有する4つのタイルに分割されてよい。すべてのタイルは、例えばデータストア２０２に書き込まれた部分的な結果を有し、次いで、次のタイルの結果に加算するために読み戻されてよい。従って、４つの書き込み及び４つの読み取りをリーチタイルに使用して、密ＤＮＮワークロードを完了してよい。データストア２０２の容量が増加するにつれて、使用される電力及びレイテンシが増大し得る。

【００２６】

４、１６、及び／又は３６の行列サイズを有するワークロードのような疎ＤＮＮワークロードの微分中、行列は、ブロックスパーシティ、疎列／行（ＣＳＣ／ＣＳＲ）の圧縮、直接インデキシング／ステップインデキシングなどのような技術を介してスパーシティ圧縮を受けて、サイズ３２の行列を得てよい。シストリックアレイシステム２００がゼロを使用して「パディング」される場合、シストリックアレイシステム２００は、第１パス上で（３６要素から）３２要素幅を有するファイル幅タイルを処理してよく、残りの４つの要素を有するタイルが続き、１６要素のタイルが続き、次いで４要素のタイルが続く。従って、シストリックアレイシステム２００の全体的な効率は、４３．７５％であってよく、これは、３２／３２＝１００％、４／３２＝１２．５％、１６／３２＝５０％、及び４／３２＝１２．５％の平均を見つけることによって計算されてよい。例えば、部分的な累積サポートを有する再構成可能なシストリックアレイを使用することにより、密ＤＮＮ並びに疎ＤＮＮワークロードの両方の処理を向上することが有益であってよい。

【００２７】

ここで図３を参照すると、図は、スケジューラ３０２を介して複数の行列乗算を処理するのに好適な部分的なバイアス累積サポート（例えば、既存のタイルストレージとは別個のバイアス累積ストレージ）を含む再構成可能なシストリックアレイ回路又はシステム３００の実施形態のブロック図である。例えば、スケジューラは、ホストプロセッサ（ＣＰＵ）、例えばプロセッサ１０２内のソフトウェアとして、ハードウェア回路として、又は再構成可能なシストリックアレイシステム３００に動作可能に接続されたそれらの組み合わせとして実装されてよい。示される実施形態では、スケジューラ３０２は、行列、例えば、タイプＡ３０４、Ｂ３０６、及び／又はＣ３０８の行列が、再構成可能なシストリックアレイシステム３００への処理のために送信される順序をスケジューリングしてよい。

【００２８】

スケジューラ３０２は、再構成可能なシストリックアレイシステム３００を介して実行するためにタイルを送信する前に、行列Ａ３０４、Ｂ３０６の特定のタイルをリオーダしてよい。また、スケジューラ３０２は、タイルをサブタイルにサイズ変更又は「ブレーク」して、再構成可能なシストリックアレイシステム３００に含まれるバイアス累積ストレージ及びロジックを利用してよい。サブタイルに分割されていないタイルは「完全」タイルと称されてよく、完全タイルの処理にはバイアス累積を使用しなくてよい。一例では、結果行列（例えば、行列Ｃ３０８）を通信するためのｘ個の読み取り／書き込みポートがある場合、スケジューラは、任意の所与の時間にｘ個より多くない完全タイルをスケジューリングしてよい。サブタイルに分割されているタイルは、「不完全」タイルと称されてよい。不完全タイルは、例えば、最後のサブタイルがスケジューリングされて、最終結果がストレージに書き出されるまで、バイアス累積ストレージ内に累積されてよい。本明細書で説明されるシステム及び方法は、以下でさらに説明されるように、完全及び不完全タイルの両方を処理し、どのタイルが完全又は不完全であるかを示し、タイルの次元を示す新しいマクロ命令を含んでよい。

【００２９】

また、本明細書に説明されるシステム及び方法は、スケジューラ３０２の出力に基づいてより小さい行列サイズを有する場合に行列データの再レイアウトをサポートしてよく、それにより、例えば、複数のＡ行列３０４をサイドバイサイドに格納及び／又は処理して

いる間に単一のＡタイルがフェッチされてよい。示される実施形態では、Ａ１及びＡ２は同一のＡ行列に属してよく、Ａ１'及びＡ２'は別のＡ行列に属してよい。アプリケーションに応じて、同一のＢ行列を複製又はコピーするか、複数のＢ行列を「繋ぎ合わせ」することのいずれかにより、Ｂタイルが形成されてよい。図示された例では、Ｂ１及びＢ１'は異なるＢ行列からのものである。しかし、Ｂ１は、レプリケートされて、それにより特定のアプリケーションに対してＢ１＝Ｂ１'となってよい。

【００３０】

特定の実施形態では、タイプＣ３０８の行列は、入力バッファから読み取られてよく、入力バッファの帯域幅は、サイクルあたりｘ読み取りに限定されてよい。従って、スケジューラ３０２は、再構成可能なシストリックアレイ３００のすべてのパスで実行するために最大Ｎ個の完全タイルをスケジューリングしてよく、従って、Ｃタイプの行列３０８の帯域幅の利用を向上する。従来の行列乗算では、Ｃ１＋＝Ａ１＊Ｂ１＋Ａ１'＊Ｂ１'＋…である。しかし、ＡとＡ'が「接着」又はともにマージされた異なる完全タイルの行列である場合、異なる演算が使用されてよい。代わりに、ハードウェア（又はソフトウェア）は、出力要素当たりより少ない演算を実行してよく、例えばＣ１＋＝Ａ１＊Ｂ１及びＣ１'＋＝Ａ２＊Ｂ１'である。しかし、通常の行列乗算より多くの出力要素があってよい。これらの追加の出力要素は、バイアス累積回路内のストレージ又はレジスタに格納されてよいか、複数の独立した宛先が使用されて、例えばスケジューラ３０２から来る「完全」及び「不完全」タイルビットに基づいてストレージに書き込んでよい。

【００３１】

前に言及されたように、本明細書で説明されるシステム及び方法は、バイアス加算累積を伴う再構成可能な行列乗算に好適な１又は複数のマクロ命令を提供してよい。新しい命令セットは、ＴＰＮＤＰＭＡＣ、『「Ｍ」累積によるタイル部分「Ｎ」ドット積』（ｔｉｌｅ ｐａｒｔｉａｌ 'Ｎ' ｄｏｔ ｐｒｏｄｕｃｔ ｗｉｔｈ 'Ｍ' ａｃｃｕｍｕｌａｔｅ）を含んでよい。ここで、Ｎはともにマージされていてよい異なる行列の数であり、Ｍは不完全な行列の数である（例えば、バイアス累積回路を使用してよい行列）。例えば、再構成可能なシストリックアレイシステム３００に入力するためのＡとしてマージされる２つの行列と、バイアス累積ロジックを使用するであろう１つと、１つのＢタイルにマージされる2つの行列との場合に、使用する命令はＴＰ２ＤＰ１ＡＣであろう。

【００３２】

一実施形態では、命令に対するフォーマットは、ＴＰＮＤＰＭＡＣ ｔｓｒｃｄｅｓｔ、ｔｓｒｃ１、ｔｓｒｃ２である。Ｎ＝１の場合に、ｔｓｒｃｄｅｓｔにより指し示される単一の行列Ｃのソース／宛先があってよい。Ｎ＞１の場合、複数のＣタイルが連続してよく、ｔｓｒｃｄｅｓｔ（例えば、ｔｓｒｃｄｅｓｔがｔｍｍ０でＮ＝２の場合はｔｍｍ０及びｔｍｍ１）で開始し、ｔｓｒｃ１が続き、次いでｔｓｒｃ２で複数のレジスタのグループを選択する。ＴＰＮＤＰＭＡＣ命令は、図４に関連して説明されるように、再構成可能なシストリックアレイシステム３００を使用して実装されてよい。

【００３３】

図４は、特定のルーティング再構成及びバイアス累積に好適な再構成可能なシストリックアレイ回路又はシステム３００の実施形態を示すブロック図である。示される実施形態では、再構成可能なシストリックアレイシステム３００の特定のコンポーネントは、シストリックアレイシステム２００に見つけられるそれらと同様に振る舞ってよい。例えば、データストレージ（例えば、複数のレジスタを有するレジスタファイル）４０２は、タイルデータのような行列タイプＡ３０４、Ｂ３０６、Ｃ３０８のデータを格納するのに使用されてよい。データストレージ４０２は、ライン４０４、４０６、４０８、及び４１０を使用して、行列Ａのタイルデータ、行列Ｂのタイルデータ、行列Ｃのタイルデータ、更新された行列Ｃのタイルデータをそれぞれ通信してよい。ライン４０４、４０６、４０８、及び４１０のそれぞれは、複数の導管を含んでよいことに留意されたい。すなわち、ライン４０４、４０６、４０８、及び４１０は、それぞれポートであってよく、各ポートは複数の導管又はラインを有してよい。ルーティング回路４１２は、行列Ａの行０及び列０に

対応する値Ａ［０］［０］を受信してよく、次いで、ルーティング回路４１２は、第１値Ａ［０］［０］を、処理要素４１４、４１６、４１８などのようなシストリックアレイシステム２００の第１行内の処理要素にブロードキャストしてよい。ルーティング回路４１２は、さらに、Ｂ内の第１行の値を表す値Ｂ［０］［０］、Ｂ［０］［１］、Ｂ［０］［２］、…、Ｂ［０］［Ｋ］を受信して、値を処理要素４１４、４１６、４１８などにブロードキャストしてよい。例えば、処理要素４１４は、値Ｂ［０］［０］を受信してよく、処理要素４１６は、値Ｂ［０］［１］を受信してよく、処理要素４１８は、値Ｂ［０］［Ｋ］を受信してよい。次いで、所与の行に対する処理要素のいくつか又はすべては、受信した入力に基づいて、乗算演算のような特定の演算の結果を出力してよい。次いで、例えば、処理要素４１４は、乗算Ａ［０］［０］＊Ｂ［０］［０］の結果を出力してよく、処理要素４１６は、乗算Ａ［０］［０］＊Ｂ［０］［１］の結果を出力してよく、処理要素４１８は、乗算Ａ［０］［０］＊Ｂ［０］［Ｋ］の結果を出力してよい。次いで、処理要素４１４、４１６、４１８の出力は、ルーティング回路４２０に送信されてよい。

【００３４】

ルーティング回路４２０は、データを処理要素４２２、４２４、４２６にルーティングしてよく、データは順番に処理要素４１４、４１６、及び４１８からカスケード「ダウン」するため、ＦＭＡ技術を適用してデータを乗算及び加算してよい。同様に、ルーティング回路４２８は、データを処理要素４３０、４３２、４３４にルーティングしてよく、データは順番に処理要素４２２、４２４、及び４２６からカスケード「ダウン」するため、ＦＭＡ技術を適用してデータを乗算及び加算してよく、ルーティング回路４３６は、データを処理要素４３８、４４０、４４２にルーティングしてよく、データは順番に処理要素４３０、４３２、及び４３４からカスケード「ダウン」するため、ＦＭＡ技術を適用してデータを乗算及び加算してよい。

【００３５】

示される実施形態は、再構成可能なルーティング回路４４４（例えば、構成スイッチを備えたルーティング回路）を含む。ルーティング回路４１２、４２０、４２８、４３６と異なり、再構成可能なルーティング回路４４４は、少なくとも２つの演算モードに基づいて異なる方法でデータをルーティングしてよい。例えば、第１の演算モードでは、再構成可能なルーティング回路４４４に含まれる構成スイッチは、オンされてよく、導出されるドット積（例えば、Ａ＊Ｂ）のチェーンの「ブレーク」が生じ、新しいチェーンを開始してよい。構成スイッチがオフにされる場合、再構成可能なシストリックアレイ３００は、１出力を有する単一のパイプラインとして振る舞ってよい。従って、値が、処理のためにパイプラインのトップ（例えば、再構成可能なシストリックアレイシステム３００の第１行）にて挿入される場合、結果は「カスケード」し、オンされる構成スイッチを有する再構成可能なルーティング回路４４４に結果が遭遇するまで下方に流れてよい。このステージにて、パイプラインは、結果値を「ブレーク」して、累積回路４４６を備えた第１バイアス加算に書き込まれてよい。結果値を対応する行列Ｃの要素に加算した後、更新値は、書き出されてよく、パイプ内の次のステージは、前の処理要素の出力値がゼロであったかのように、ロードされる。従って、オンされる構成スイッチを有する再構成可能なルーティング回路４４４とのカスケーディング値の遭遇は、新しいパイプラインの開始と見なされてよい。複数の再構成可能なルーティング回路４４４が使用されてよく、例えば、再構成可能なルーティング回路４４４は、第４行ごとなど、８行の再構成可能なシストリックアレイシステム３００内に配置されてよく、従って、複数の再構成可能なルーティング回路４４４が使用されてよいことを理解されたい。

【００３６】

一実施形態では、第１の演算モードにある場合、データストア４０２の第１の複数のレジスタに格納された値は、単一入力の２次元行列Ａを表してよく、データストア４０２の第２の複数のレジスタに格納された値は、単一入力の２次元行列Ｂを表してよく、一方、データストア４０２の第３の複数のレジスタに格納された値は、単一入力の２次元行列Ｃを表してよい。第２の演算モードにある場合、データストア４０２の第１の複数のレジス

タに格納された値は、複数入力の２次元行列Ａ及びＡ'を表してよく、データストア４０２の第２の複数のレジスタに格納された値は、複数入力の２次元行列Ｂ及びＢ'を表してよく、一方、データストア４０２の第３の複数のレジスタに格納された値は、複数入力の２次元行列Ｃ及びＣ'を表してよい。

【００３７】

特定の実施形態では、第１の演算モードの実行中、再構成可能なシストリックアレイシステム３００は、タイルＡ及びタイルＢからそれぞれのルーティング回路に値を送信してよい。例えば、演算は、第１の演算モードの場合に、タイルＡからの行列Ａにタイルからの行列Ｂを乗算し、次いで、タイルＣからの行列Ｃの対応する値にそれぞれの結果を加算し、第２の演算モードの場合に、タイルＡからの行列Ａにタイルからの行列Ｂを乗算し、次いで、タイルＣからの行列Ｃの対応する値にそれぞれの結果を加算するとともに、タイルＡからの行列Ａ'にタイルＢからの行列Ｂ'を乗算し、次いで、タイルＣからの行列Ｃ'の対応する値にそれぞれの結果を加算してよい。第１の演算モードでは、処理要素４３８、４４０、４４２の出力は、累積回路４４６を備えた第１バイアス加算をバイパスしてよく、処理要素４４８、４５０、４５２に直接提供されてよい。次いで、処理要素４４８、４５０、４５２は、上述のような乗算及び加算を適用し、次いで、それぞれの出力を累積回路４５４を備えた第２バイアス加算に提供してよい。次いで、累積回路４５４を備えた第２バイアス加算は、処理要素４４８、４５０、４５２から提供された出力を使用して、行列Ｃを更新してよい。

【００３８】

第２の演算モードでは、処理要素４３８、４４０、４４２の出力は、累積回路４４６を備えた第１バイアス加算により使用されて、例えば、特定の値を加算し格納してよい。前に言及したように、再構成可能なルーティング回路４４４がオンされた構成スイッチを有する場合、再構成可能なルーティング回路４４４は、入力として提供される値を乗算及び加算し、結果を更新行列Ｃに送信してよいが、結果（例えば、乗算及び加算の結果）を後の微分で使用するために累積してもよい。第２の演算モードでは、処理要素４４８、４５０、４５２は、処理要素４３８、４４０、４４２の出力に代えてゼロを受信してよく、従って、処理要素４４８、４５０、４５２にて開始する演算は新しいパイプラインとして進んでよい。累積回路４５４を備えた第２バイアス加算は、第１の演算モード（例えば、値の累積をバイパスする）を提供するためにスイッチオフされる、又は第２の演算モードに対してスイッチオンされる累積スイッチを有してよい。

【００３９】

図５は、再構成可能なルーティング回路４４４及び累積回路５０１を備えたバイアス加算（例えば、回路４４６又は４５４と同等）の実施形態を示す概略図である。示される実施形態では、再構成可能なシストリックアレイシステム３００の処理要素の行（例えば、行３）からのデータ５００は、再構成可能なシストリックアレイシステム３００の下流行５０２（例えば、行４）に提供されてよい。処理要素の下流行５０２は、行列Ｂのデータ５０４及び行列Ａのデータ５０６を受信してもよい。次いで、行５０２内の処理要素は、例えばライン５０８を介して再構成可能なルーティング回路４４４への出力を提供してよい。

【００４０】

再構成可能なルーティング回路４４４は、デマルチプレクサ５１０及びマルチプレクサ５１２を含んでよく、それにより、デマルチプレクサ５１０及びマルチプレクサ５１２の両方がスイッチとして使用される。すなわち、デマルチプレクサ５１０及びマルチプレクサ５１２は、同一の信号（例えば、構成オン又はオフ信号）を受信し、ともにデータルーティングのスイッチとして機能してよい。再構成可能なルーティング回路４４４がデマルチプレクサ５１０及びマルチプレクサ５１２へのセレクタを介してオンされる場合、デマルチプレクサ５１０は行５０２の処理要素を介して導出される出力をライン５１４を介して累積回路５０１を備えたバイアス加算に書き込んでよい。順番に、マルチプレクサ５１２は、ゼロを下流行５１６（例えば、行５）の処理要素に、例えばライン５１８を介して

送信してよい。従って、行５１６の処理要素は、行５０２からのデータを使用しなくてよく、代わりに、行列Ｂのデータ５０４及び行列Ａのデータ５０６を使用して出力５２０を導出してよく、次いで、その出力は次の下流行（例えば、行６）送信されてよい。

【００４１】

累積イネーブル信号５２２がオンされる場合、累積回路５０１を備えたバイアス加算は、バイアス５２４をＣタイル５２６に加算するとともに結果を格納そうでなければ累積してよい。累積イネーブル信号５２２は、アドレスチェック信号５３２によって受信される累積イネーブル信号５３０（例えば、マクロ命令の実行に基づく信号）のＢｏｏｌｅａｎ ＯＲを導出するＯＲゲート５２８を使用することによりオンされてよい。アドレスチェック信号５３２は、行列Ｃのタイルアドレス５３４の表示であってよい。より具体的には、マイクロアーキテクチャのサポートが提供されてよく、それにより、ハードウェア内でＣタイルアドレス５３４がチェックされて、宛先の衝突が発生するか、例えば２つの行列演算は同一の行列Ｃの宛先アドレスを共有するか、について判断する。次いで、アドレスが同一の場合、累積ロジックは自動的にオンされて、例えば、宛先の上書きを防止する。最後のサブタイルビット５３６が受信されると（例えば、スケジューラ３０２から到着すると）、累積回路５０１を備えたバイアス加算は、例えば、すべてのレジスタに亘って、すべての累積値を加算してよい。すなわち、最後のサブタイルビット５３６は、すべてのサブタイルがここで送信されていて、従って、任意の累積値がここで累積回路５０１を備えたバイアス加算を介して加算及び格納されてよいことを示してよい。

【００４２】

再構成可能なルーティング回路４４４が（例えば、デマルチプレクサ５１０及びマルチプレクサ５１２へのセレクタを介して）オフにされる場合、デマルチプレクサ５１０は、行５０２の処理要素によりライン５３５を介してマルチプレクサ５１２に導出される出力を送信してよい。次いで、マルチプレクサ５１２は、行５０２の処理要素により導出される出力をライン５１８を介して下流行５１６に送信してもよい。従って、再構成可能なルーティング回路４４４をオフすると、再構成可能なルーティング回路４４４が、行５０２の処理要素及び行５１６の処理要素の間のパススルースイッチとして機能するようになってよい。再構成可能なルーティング回路４４４を提供することにより、本明細書で説明される技術は、再構成可能なシストリックアレイシステム３００を通じてデータのより効率的なルーティングを可能にしてよい。

【００４３】

図６は、さらに詳細に示す累積回路５０１を備えたバイアス加算の実施形態の概略図である。累積回路５０１を備えたバイアス加算は、メモリストレージの一致するステージを使用することにより、特定のレイテンシ（例えば、加算器のレイテンシ）を説明するように設計されてよい。例えば、３つのステージが使用されて３のレイテンシを一致させてよく、４つのステージが使用されて４のレイテンシ（例えば、2^3+1）などを一致させてよい。示される実施形態では、カウンタ６００は、レイテンシに基づいてカウントするために使用されてよい。従って、３ビットのカウンタは３のレイテンシに使用されてよく、４ビットのカウンタは７より大きいレイテンシなどに使用されてよい。従って、適切な値を多重選択するためのサイズが、増加してもよい。累積回路５０１を備えた加算への入力は、ドット積６０２（例えば、再構成可能なシストリックアレイ３００の行からのＡ＊Ｂ）、加算するバイアス６０４（例えば、図４に示されるライン４０８を介した行列Ｃ）、クロック信号６０６（例えば、再構成可能なシストリックアレイ３００のクロック）、最後のサブタイル信号５３６（図５にも示される）、及び累積イネーブル信号５２２（図５にも示される）を含んでよい。

【００４４】

演算中、加算器６０８は、ドット積６０２をマルチプレクサ６１０からの出力と加算してよい。マルチプレクサ６１０の出力は、ＡＮＤゲート６１２からの信号を介して選択されてよい。ＡＮＤゲート６１２は、カウントリセット信号６１４及びクロックフリップフロップ６１６からの出力の間のＢｏｏｌｅａｎ ＡＮＤを実行してよい。クロックフリッ

プフロップ６１６は、ＡＮＤゲート６１８から出力されるデータを格納してよい。例えば、ＡＮＤゲート６１８は、カウンタ６００の出力及び最後のサブタイル信号５３６の間のＢｏｏｌｅａｎ　ＡＮＤ演算を実行してよい。累積イネーブル信号５２２がオンの場合、カウンタ６００は、例えば、デマルチプレクサ６２５を選択することによりクロック信号６０６が送信されると、記憶回路（例えば、フリップフロップのような記憶コンポーネント）６２０、６２２、６２４へのドット積６０２のストレージを指示してよい。

【００４５】

次いで、最後のサブタイル信号５３６は、ストレージ６２０、６２２、６２４が、ＡＮＤゲート６２６、６２８、６３０を通じて累積されたデータ値をパスし、加算器６３２、６３４を介して加算されるようにしてよい。次いで、累積イネーブル信号５２２を使用して、累積されたデータ値がマルチプレクサ６３６の出力として選択されてよい。加算器６３８は、使用されて、累積されたデータ値をマルチプレクサ６４２の加算バイアスセレクタ信号を介して行列Ｃのバイアス６０４に加算してよい。行列Ｃのバイアス６０４は、ストレージ６４３から到着してよい。次いで、加算の結果は、例えばライン４１０（図４にも示される）を介して、更新された行列Ｃに提供されてよい。前に言及したように、累積回路５０１を備えたバイアス加算は、特定のレイテンシを考慮して設計されてよい。図示された例では、３つのストレージ６２０、６２２、６２４は、３又はより小さいレイテンシを処理してよい。しかし、演算中にレイテンシが増加し得る場合がある。レイテンシが増加した場合、ライン６４４が使用されて、例えば加算器６０８を介して、ストレージ６２０、６２２、６２４に格納されているより古い値で新しいドット積６０２を加算することにより、ループ内の値を連続的に累積してよい。累積イネーブル信号５２２がオフにされる場合、ドット積６０２は、デマルチプレクサ６４６をトラバースし、次いでマルチプレクサ６３６をトラバースし、その後、加算器６３８により行列Ｃのバイアス６０４に加算されてよい。累積回路５０１を備えたバイアス加算に提供することにより、本明細書で説明された技術は、密及び疎ＤＮＮの両方をより効率的に処理するとともに、より柔軟なシストリックアレイベースの計算を提供してよい。

【００４６】

前に言及したように、複数の再構成可能なルーティング回路４４４が使用されてよい。同様に、累積回路を備えた複数のバイアス加算、例えば、累積回路５０１を備えたバイアス加算が提供されてよい。ここで図７を参照すると、図は、複数のルーティング回路及び累積回路を備えた複数のバイアス加算（例えば、部分的なバイアス累積サポート）を含む再構成可能なシストリックアレイシステム７００を示すブロック図である。示される実施形態では、シストリックアレイシステム７００は、データストレージ７０２（例えば、複数のレジスタを有するレジスタファイル）を含む。データストレージ７０２は、ライン７０４、７０６、７０８、及び７１０を使用して、行列Ａのタイルデータ、行列Ｂのタイルデータ、行列Ｃのタイルデータ、更新された行列Ｃのタイルデータをそれぞれ通信してよい。ライン７０４、７０６、７０８、及び７１０のそれぞれは、複数の導管を含んでよいことに留意されたい。すなわち、ライン７０４、７０６、７０８、及び７１０は、それぞれポートであってよく、各ポートは複数の導管又はラインを有してよい。

【００４７】

示される実施形態はまた、８つの回路ブロック７１２、７１４、７１６、７１８、７２０、７２２、７２４、７２６を含む。回路ブロック７１２、７１４、７１６、７１８、７２０、７２２、７２４、７２６のそれぞれは、１又は複数行の処理要素を含んでよく、ここで、処理要素は融合型乗算加算ユニット（ＦＭＡ）を含んでよい。一実施形態では、再構成可能なシストリックアレイシステム７００が３２行の処理要素を有する場合のように、回路ブロック７１２、７１４、７１６、７１８、７２０、７２２、７２４、７２６のそれぞれは４行の処理要素を含んでよい。データが第１回路ブロック７１２に入ると、データは、カスケード方式で処理されてよく、その後、回路ブロック７１４、７１６、７１８、７２０、７２２、７２４、及び７２６を通じてカスケード順序で流れ、例えばＣ＋＝Ａ＊Ｂを計算してよい。

【００４８】
図示のとおり、再構成可能なルーティング回路７２８、７３０、７３２、７３４、７３６、７３８、７４０は、回路ブロック７１２、７１４、７１６、７１８、７２０、７２２、７２４の下流に配置されてよい。再構成可能なルーティング回路７２８、７３０、７３２、７３４、７３６、７３８、７４０のそれぞれは、例えばスイッチングを介して、累積回路７４２、７４４、７４６、７４８、７５０、７５２、７５４を備えた下流のバイアス加算へのデータのフローを可能にしてよい。再構成可能なルーティング回路７２８、７３０、７３２、７３４、７３６、７３８、７４０は、さらに、例えば前述したようにスイッチオンされる場合に「新」しいパイプラインの作成を可能にしてよい。累積回路７４２、７４４、７４６、７４８、７５０および７５２を備えたバイアス加算のそれぞれは、上の図６に説明したように、ドット積へのバイアスを加算及び累積値をバイアスに加算するのに好適であってよい。ルーティング回路７５７は、スイッチング機能を含まなくてよく、従って、回路ブロック７２６からの出力値を、累積しないでバイアス加算のために直接パスすることによって、データをバイアス加算回路７５６に送信してよい。従って、再構成可能なシストリックアレイシステム７００は、より効率的に且つ柔軟に、Ｃ＋＝Ａ＊Ｂを含む様々な計算を導出してよい。

【００４９】
本明細書で説明された技術をプログラムで使用するために、特定の命令（例えば、マクロ命令）が提供される。例えば、ＴＰＮＤＰＭＡＣは、ＴＰ２ＤＰ１ＡＣのような命令又は「１が累積されたタイル部分の２ドット積」（ｔｉｌｅ　ｐａｒｔｉａｌ　２　ｄｏｔ　ｐｒｏｄｕｃｔ　ｗｉｔｈ　１　ａｃｃｕｍｕｌａｔｅ）のプログラムによる使用をもたらしてよい。ＴＰ２ＤＰ１ＡＣ命令は、処理要素のアレイの中央にある再構成可能なルーティング回路（例えば、再構成可能なルーティング回路７３４）をオンすることにより、また累積回路を備えた対応するバイアス加算（例えば、累積回路７４８を備えたバイアス加算）をスイッチオンすることにより、ともにマージされた２つの均一サイズの行列を処理してよい。

【００５０】
異なるサイズの行列がともにマージされる場合、ＴＳＺＤＰ「ドット積のタイルサイズ」（ｔｉｌｅ　ｓｉｚｅｓ　ｆｏｒ　ｄｏｔ　ｐｒｏｄｕｃｔｓ）マクロ命令が使用されてよい。一実施形態では、ＴＳＺＤＰマクロ命令は、Ａ、Ｂ、及びＣレジスタオペランドに加えて、ともにマージされる行列のサイズを指定する即値を取ってよい。別の実施形態では、サイズはエンコードされてよい。例えば、４の倍数（例えば、最大３２）を有する行列のようなマージング行列がサポートされる場合、様々な行列サイズを次のようにエンコードしてよい。

【００５１】

【表１】

即時エンコーディング	第１Ｋサイズ(K1)	第２Ｋサイズ(K2)	第３Ｋサイズ(K3)	第４Ｋサイズ(K4)	第５Ｋサイズ(K5)	第６Ｋサイズ(K6)	第７Ｋサイズ(K7)	第８Ｋサイズ(K8)	デコーダ出力	構成スイッチ
0000000	32	0	0	0	0	0	0	0	0000000	すべてのスイッチがダウン
0000001	4	28	0	0	0	0	0	0	0000001	スイッチ1がオン
0000010	8	24	0	0	0	0	0	0	0000010	スイッチ２がオン
0000011	4	4	24	0	0	0	0	0	0000011	スイッチ1及び2がオン
0000100	12	20	0	0	0	0	0	0	0000100	スイッチ３がオン
. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .
1111110	8	4	4	4	4	4	4	0	1111110	スイッチ1を除くすべてのスイッチがオン
1111111	4	4	4	4	4	4	4	4	1111111	すべてのスイッチがオン

【００５２】

表１は、スイッチの使用に言及し、スイッチは、順番に、図７の同等の再構成可能なルーティング回路の使用に言及する。例えば、スイッチ１は、再構成可能なルーティング回路７２８を指してよく、スイッチ２は、再構成可能なルーティング回路７３０、再構成可能なルーティング回路７３２などを指してよい。即時エンコーディング値がゼロの場合、これはタイルサイズが３２であり、Ａ及びＢ入力の両方の入力として単一の行列が使用されることを意味する。即時エンコーディング値１１１１１１０は、第１のスイッチを除くすべての構成スイッチを有効にしてよく、それにより、完全なシストリックアレイ７００は７つの独立した小さなアレイと見なしてよく、第１のアレイは８の行列サイズを処理することが可能であり、他のすべてはそれぞれ４の行列サイズを処理してよい。同様に、１１１１１１１の即時エンコーディング値は、すべての構成スイッチを有効にしてよく、それにより、完全なシストリックアレイ回路７００は、各回路が４の行列サイズを処理可能である小さなアレイの８つの独立した回路と見なされてよい。

【００５３】

本明細書ではＴＡＣＤＰ「タイル累積ドット積」（ｔｉｌｅ　ａｃｃｕｍｕｌａｔｅ　ｄｏｔ　ｐｒｏｄｕｃｔ）と称される、構成スイッチを制御する前述の命令に基づいて累積ロジックを有効及び無効にする命令が使用されてもよい。この命令は、適切な構成スイッチ値でのみ有効であってよいことに留意されたい（すなわち、構成スイッチが有効でない場合、パイプの最後にあるルーティング回路７５７の終わりの構成スイッチを除いて、累積ロジックは有効でなくてよく、構成スイッチを含まなくてよいため、再構成可能でなくてよい）。アキュムレータは、ＴＡＣＤＰ　ｉｍｍ＿ａｃ＃形式の即値をパスすることにより又は命令を通じてパスされる即値を介して有効にされてよい（例えば、ＴＰ２ＤＰ　ｔｓｒｃｄｅｓｔ、ｔｓｒｃ１、ｔｓｒｃ２、ｉｍｍ＿ｓｚ＃、ｉｍｍ＿ａｃ＃）。このＴＡＣＤＰ命令は、ＴＳＺＤＰ命令（ＴＳＺＤＰ　ｉｍｍ＿ｓｚ＃、ｉｍｍ＿ａｃ＃）とマージされてもよい。ＴＡＣＤＰ即時エンコーディングは次のとおりでよい。

【表２】

即時エンコーディング	蓄積スイッチ1	蓄積スイッチ2	蓄積スイッチ3	蓄積スイッチ4	蓄積スイッチ5	蓄積スイッチ6	蓄積スイッチ7	蓄積スイッチ8
0000000	0	0	0	0	0	0	0	0
0000001	1	0	0	0	0	0	0	0
0000010	0	1	0	0	0	0	0	0
0000011	1	1	0	0	0	0	0	0
0000100	0	0	1	0	0	0	0	0
. .	. .	. .	. .	. .	. .	. .	. .	. .
1111110	0	1	1	1	1	1	1	1
1111111	1	1	1	1	1	1	1	1

【００５４】

累積回路（例えば、回路７４２、７４４、７４６、７４８、７５０、７５２）を備えたバイアス加算は、累積スイッチ（例えば、累積イネーブル信号５２２）がオンされる場合にのみ有効になってよく、そうでない場合、特定のデータが再構成可能なシストリックアレイ７００のこのセクション（例えば、累積回路を備えたバイアス加算）に入らなくてよい。しかし、累積イネーブル信号を有さなくてよいバイアス加算回路７５６が使用される。すなわち、ルーティング回路７５７は、データをバイアス加算回路７５６のみに直接ルーティングしてよく、スイッチング機能を提供しなくてよい。いくつかの実施形態では、累積回路７５２を備えた最後のバイアス加算は、累積が常にスイッチオンされると仮定して動作されてよい。

【００５５】

累積ロジックは、マイクロアーキテクチャモード及びアーキテクチャモードの２つのモードにより有効にされてよい。マイクロアーキテクチャモードでは、例えば図５に示されるアドレスチェック５３２を介して、前の宛先（例えば、タイルレジスタｔｍｍ０）及び現在の宛先アドレスが同一であると識別された場合、再構成可能なシストリックアレイ７００及び関連ハードウェアは、累積を可能にしてよい。アーキテクチャモードでは、累積は、ロジックがＴＳＺＤＰ命令又はＴＡＣＤＰのいずれかにより制御される命令により有効にされてよい。前に言及したように、最後のグループの処理要素行（例えば、図示の例における回路ブロック７２６）に関連する構成スイッチがないため、構成スイッチなしで最後のバイアス加算回路７５６のみがオンされてよい。

【００５６】

図８は、本明細書で説明される技術を実装するために使用されてよい処理８００の実施形態を示す。処理８００は、再構成可能なシストリックアレイ３００、７００及びマクロ命令ＴＰＮＤＰＭＡＣ、ＴＳＺＤＰ及び／又はＴＡＣＤＰを介してのようなハードウェア及び／又はソフトウェアとして実装されてよい。示される実施形態では、解かれる問題のタイルサイズを決定してよい（ブロック８０２）。問題は、密ＤＮＮ、疎ＤＮＮ、又はそれらの組み合わせとともに、機械学習、ビデオ処理、音声認識、画像認識、データ圧縮、データベース検索ランキング、バイオインフォマティクス、ネットワークセキュリティパターンの認識、空間ナビゲーションなどにおける問題を含んでよい。例えば、使用されるシストリックアレイの行及び列の数（例えば、再構成可能なシストリックアレイシステム３００、７００）に基づいて、タイルサイズを選択して、例えば加算されたゼロを最小化することによりアレイをより快適に合致させてよい。タイルサイズが選択されると（ブロック８０２）、いくつかの完全及び／又は不完全タイルが導出されてよい（ブロック８０４）。完全タイル８０６は、それら全体において使用されるシストリックアレイに合致してよいが、不完全タイル８０８はサブタイルに細分化されてよい。

【００５７】

次いで、完全タイル８０６及び不完全タイル８０８が処理されてよい（ブロック８１０）。例えば、例えば図５に示されるアドレスチェック５３２を介して、前の宛先（例えば

、タイルレジスタｔｍｍ０）及び現在の宛先アドレスが同一であると識別された場合、マイクロアーキテクチャモードが使用されて、使用されるシストリックアレイを実行し、自動的に宛先の衝突を検出し、累積ロジックをスイッチオンしてよい。アーキテクチャモードでは、累積は、ＴＳＺＤＰ命令及び／又はＴＡＣＤＰにより有効にされてよい。次いで、計算の結果が提供されてよい（ブロック８１２）。例えば、Ｃ＋＝Ａ＊Ｂに基づく最終Ｃは、計算された行列Ｃのそれぞれに提供されてよい。本明細書で説明される回路（例えば、再構成可能なシストリックアレイシステム３００、７００）は、ハードウェアアクセラレータの一部として、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）として、特定用途向け集積回路（ＡＳＩＣ）として、カスタムマイクロチップとして、又はそれらの組み合わせとしてマイクロプロセッサに実装されてよいことを理解されたい。

【００５８】

本開示に記載される実施形態が様々な修正及び代替的形態に影響されやすい場合がある一方で、具体的な実施形態が、図面における例により示されており、本明細書において詳細に説明されている。しかしながら、本開示は開示された特定の形態に限定されることが意図されているものではないことが理解され得る。本開示は、以下に添付される特許請求の範囲により定義されるように、本開示の趣旨及び範囲内に含まれるすべての修正、均等物、および代替物をカバーするものである。

【００５９】

本明細書によれば、以下の各項目に記載の構成もまた開示される。

［項目１］

データを格納するように構成されたデータストレージと、

再構成可能なシストリックアレイ回路と、を備え、前記再構成可能なシストリックアレイ回路は、

前記データを処理するように構成された１又は複数のグループの処理要素を有する第１回路ブロックと、

前記データを処理するように構成された１又は複数のグループの処理要素を有する第２回路ブロックと、

行列バイアスを累積値に、乗算積に、又はそれらの組み合わせに加算するように構成された累積回路を備えた第１バイアス加算と、

微分を前記第１回路ブロックから前記第２回路ブロックに、前記第１回路ブロックから累積回路を備えた前記第１バイアス加算に、又はそれらの組み合わせにルーティングするように構成された第１ルーティング回路と、を含む、システム。

［項目２］

前記第１ルーティング回路は、構成スイッチ信号を受信することに基づいて、前記微分を前記第１回路ブロックから前記第２回路ブロックに、前記第１回路ブロックから累積回路を備えた前記第１バイアス加算に、又はそれらの組み合わせにルーティングするように構成されたデマルチプレクサ及びマルチプレクサ回路を含む、項目１に記載のシステム。

［項目３］

累積回路を備えた前記第１バイアス加算は、クロック信号に基づいて前記乗算積を前記累積値として累積するように構成された記憶回路、及び前記行列バイアスを前記累積値に、前記乗算積に、又はそれらの組み合わせに加算するように構成された少なくとも１つの加算器を含む、項目１に記載のシステム。

［項目４］

累積回路を備えた前記第１バイアス加算は、Ｎの加算器のレイテンシを含み、前記記憶回路は、Ｎ個の記憶コンポーネントを含む、項目３に記載のシステム。

［項目５］

前記Ｎ個の記憶コンポーネントは、それぞれ、フリップフロップを含む、項目４に記載のシステム。

［項目６］

前記記憶回路は、前記Ｎ個の記憶コンポーネントをマルチプレクサに結合するＮ個のラ

インを含み、前記記憶回路は、前記加算器のレイテンシが演算中にＮを超えた場合に、累積値を前記Ｎ個の記憶コンポーネントから前記マルチプレクサに前記Ｎ個のラインを介して転送するように構成される、項目４に記載のシステム。

［項目７］

累積回路を備えた前記第１バイアス加算は、累積回路を備えた前記第１バイアス加算に入る新しい値を前記累積値に加算し、前記加算の結果を前記Ｎ個の記憶コンポーネントに格納するように構成される、項目６に記載のシステム。

［項目８］

１又は複数のグループの処理要素を有する第３回路ブロックと、

第２行列バイアスを第２累積値に、前記乗算積に、又はそれらの組み合わせに加算するように構成された累積回路を備えた第２バイアス加算と、

微分を前記第２回路ブロックから前記第３回路ブロックに、前記第２回路ブロックから累積回路を備えた前記第２バイアス加算に、又はそれらの組み合わせにルーティングするように構成された第２ルーティング回路と、を備える、項目１に記載のシステム。

［項目９］

前記第３回路ブロックの下流に配置され、第３行列バイアスを前記第３回路ブロックから出力に加算するように構成されたバイアス加算回路を備える、項目８に記載のシステム。

［項目１０］

前記再構成可能なシストリックアレイ回路を使用するよう、又は前記再構成可能なシストリックアレイ回路を含むように構成されたホストプロセッサ（ＣＰＵ）を備え、

前記ＣＰＵは、『「Ｍ」累積によるタイル部分「Ｎ」ドット積』命令であり、前記Ｎは、ともにマージされている異なる行列の数であり、Ｍは、前記再構成可能なシストリックアレイ回路への入力として使用される不完全な行列の数である、前記命令、前記再構成可能なシストリックアレイ回路への入力として使用されるとともにマージされている前記異なる行列のサイズを指定する即値を有する「ドット積に対するタイルサイズ」命令、累積回路を備えた前記第１バイアス加算を制御する「タイル累積ドット積」命令、又はそれらの組み合わせを実行するように構成される、項目１に記載のシステム。

［項目１１］

行列Ａ及び行列Ｂに基づいてデータの１又は複数のタイルのそれぞれに対してタイルサイズを決定する段階と、

タイルサイズに基づいて、完全タイル、不完全タイル、又はそれらの組み合わせを導出する段階と、

行列Ｃの結果を導出するために、前記完全タイル、前記不完全タイル、又はそれらの組み合わせを、再構成可能なシストリックアレイ回路を介して処理する段階であり、前記完全タイル、前記不完全タイル、又はそれらの組み合わせを処理することは、前記行列Ｃの結果を提供するために、前記再構成可能なシストリックアレイ回路に含まれるルーティング回路及び前記再構成可能なシストリックアレイ回路に含まれる累積回路を備えたバイアス加算又はそれらの組み合わせを適用することを含む、段階と、を備える方法。

［項目１２］

前記再構成可能なシストリックアレイ回路は、Ｎ行Ｍ列のアレイサイズを含み、前記完全タイルは、Ｎ行以下及びＭ列以下を有する完全サイズを含み、前記不完全タイルは、Ｎ行より多く、Ｍ列より多く、又はそれらの組み合わせを有する不完全サイズを含む、項目１１に記載の方法。

［項目１３］

前記ルーティング回路を適用する段階は、微分を１又は複数のグループの処理要素を含む第１回路ブロックから１又は複数のグループの処理要素を含む第２回路ブロックにルーティングする段階、及び微分を前記第１回路ブロックから累積回路を備えた前記バイアス加算に、又はそれらの組み合わせにルーティングする段階を含む、項目１１に記載の方法。

［項目１４］　微分を前記第１回路ブロックから累積回路を備えた前記バイアス加算にルーティングする段階は、前記微分を累積回路を備えた前記バイアス加算にて受信し、前記微分を行列Ｃのバイアスに加算するために累積値に累積する段階を含む、項目１３に記載の方法。

［項目１５］

前記完全タイル、前記不完全タイル、又はそれらの組み合わせを前記再構成可能なシストリックアレイ回路を介して処理する段階は、

行列Ｃのアドレス衝突を検出し、累積回路を備えた前記バイアス加算に通信される累積イネーブル信号を自動的にオンにするように構成されたマイクロアーキテクチャモードを適用する段階と、

前記再構成可能なシストリックアレイ回路への入力として使用されるとともにマージされている前記異なる行列のサイズを指定する即値を有する「ドット積に対するタイルサイズ」命令、累積回路を備えた前記バイアス加算を制御する「タイル累積ドット積」命令、又はそれらの組み合わせを実行することによりアーキテクチャモードを適用する段階と、を含む、項目１１に記載の方法。

［項目１６］

データを格納するように構成されたデータストレージと、

再構成可能なシストリックアレイ回路と、

前記再構成可能なシストリックアレイ回路に接続されるコアのデコーダであり、単一の命令をデコードされた１又は複数の命令にデコードする、デコーダと、を備え、

前記１又は複数の命令は、

行列Ａ及び行列Ｂを表す前記データを、前記データストレージから、前記データを処理し、前記データに基づく微分を提供するように構成された１又は複数のグループの処理要素を含む第１回路ブロックに通信し、

再構成可能なルーティング回路のスイッチングオン又はオフに基づいて、前記微分を前記第１回路ブロックから第２回路ブロックに、累積回路を備えたバイアス加算に、又はそれらの組み合わせにルーティングするように構成され、

累積回路を備えた前記バイアス加算は、行列バイアスを、累積値に、行列Ａ及び行列Ｂの乗算積に、又はそれらの組み合わせに加算するように構成され、

前記第１回路ブロック、前記第２回路ブロック、前記再構成可能なルーティング回路、累積回路を備えた前記バイアス加算、又はそれらの組み合わせは、前記再構成可能なシストリックアレイ回路に含まれる、装置。

［項目１７］

前記単一の命令は、デコードされると、前記再構成可能なシストリックアレイ回路への入力として使用されるとともにマージされている異なる行列のサイズを指定する即値を有する「ドット積に対するタイルサイズ」命令、累積回路を備えた前記バイアス加算を制御する「タイル累積ドット積」命令、又はそれらの組み合わせを介してアーキテクチャモードを使用する、項目１６に記載の装置。

［項目１８］

前記単一の命令は、『「Ｍ」累積によるタイル部分「Ｎ」ドット積』命令を含み、Ｎは、ともにマージされている異なる行列の数であり、Ｍは、前記再構成可能なシストリックアレイ回路への入力として使用される不完全な行列の数である、項目１７に記載の装置。

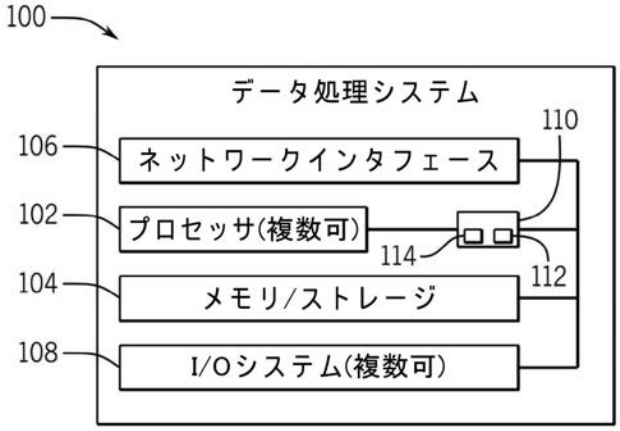
［項目１９］

前記単一の命令は、デコードされると、前記再構成可能なシストリックアレイ回路に、前記データを使用することによってＣ＝＋Ａ＊Ｂを解かせ、前記データは、前記行列Ａ及び前記行列Ｂを表す、項目１６に記載の装置。

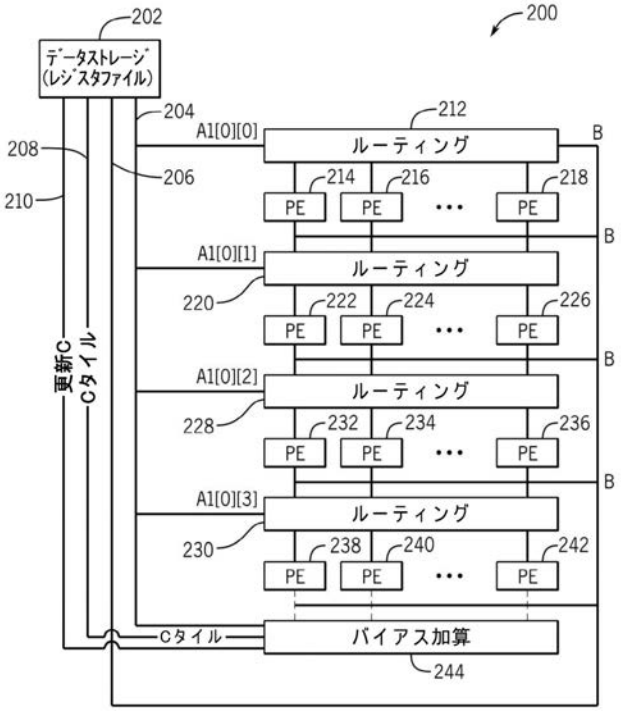
［項目２０］　前記再構成可能なシストリックアレイ回路を有する回路を備え、

前記回路は、マイクロプロセッサ、ハードウェアアクセラレータ、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）、特定用途向け集積回路（ＡＳＩＣ）、カスタムマイクロチップ、又はそれらの組み合わせを含む、項目１６に記載の装置。

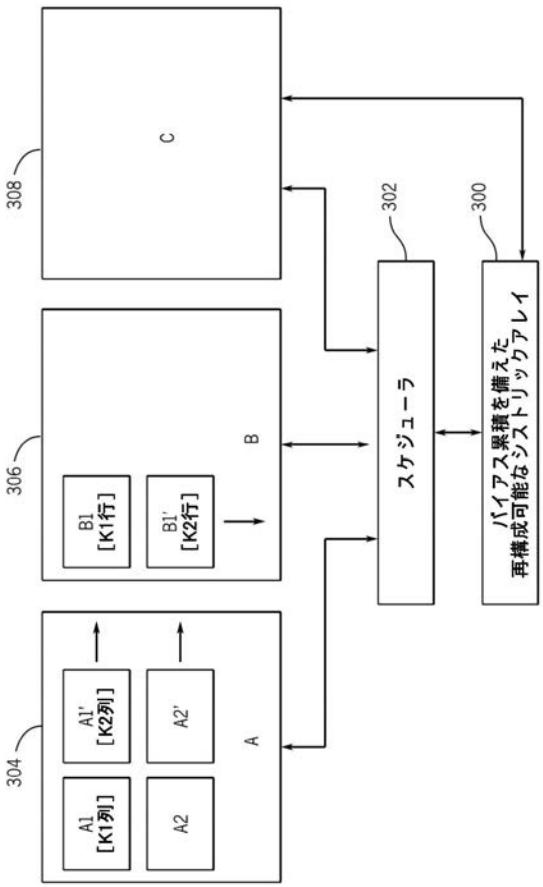
【図１】


100
データ処理システム
106
ネットワークインタフェース
102
プロセッサ(複数可)
104
メモリ/ストレージ
108
I/Oシステム(複数可)
110
112
114

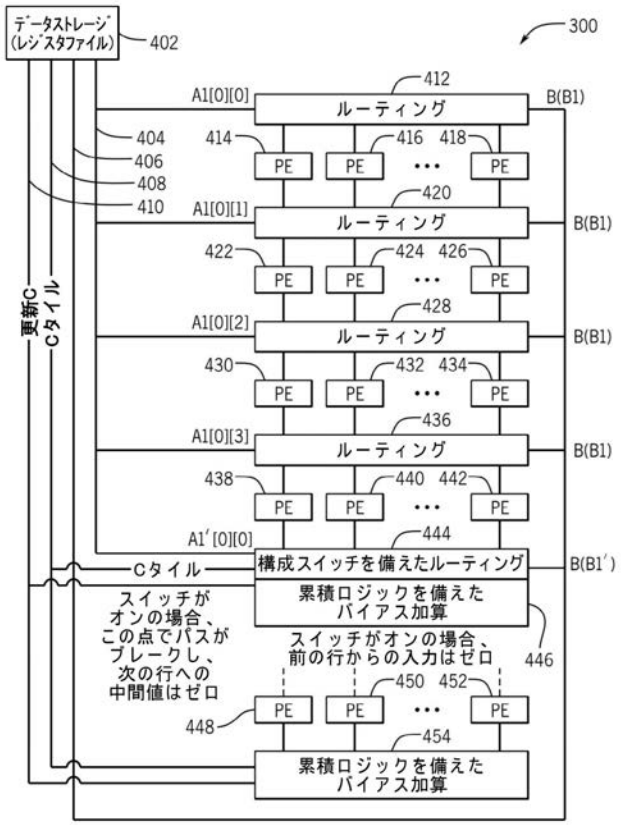
【図２】


200
202
データストレージ
(レジスタファイル)
204
206
208
210
212
A1[0][0]
ルーティング
B
214
216
218
PE
A1[0][1]
220
222
224
226
A1[0][2]
228
232
234
236
A1[0][3]
230
238
240
242
更新C
Cタイル
Cタイル
バイアス加算
244

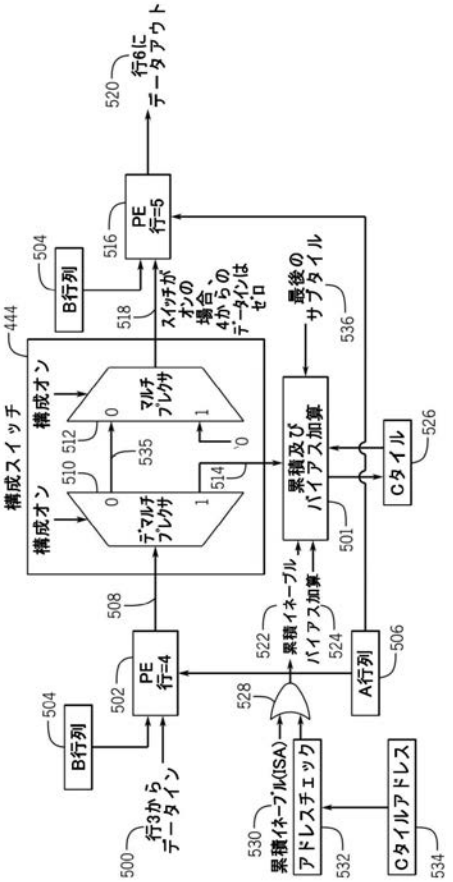
【図３】


C
308
B
306
B1
[K1行]
B1'
[K2行]
A
304
A1'
[K2列]
A2'
A1
[K1列]
A2
スケジューラ
302
バイアス累積を備えた
再構成可能なシストリックアレイ
300

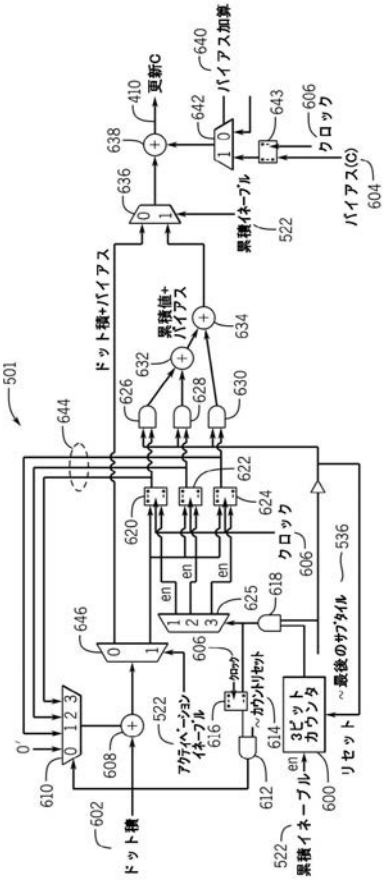
【図４】


300
データストレージ
(レジスタファイル)
402
404
406
408
410
412
A1[0][0]
ルーティング
B(B1)
414
416
418
PE
A1[0][1]
420
422
424
426
A1[0][2]
428
430
432
434
A1[0][3]
436
438
440
442
A1'[0][0]
444
構成スイッチを備えたルーティング
B(B1')
Cタイル
更新C
Cタイル
累積ロジックを備えた
バイアス加算
446
スイッチが
オンの場合、
この点でパスが
ブレークし、
次の行への
中間値はゼロ
スイッチがオンの場合、
前の行からの入力はゼロ
448
450
452
454
累積ロジックを備えた
バイアス加算

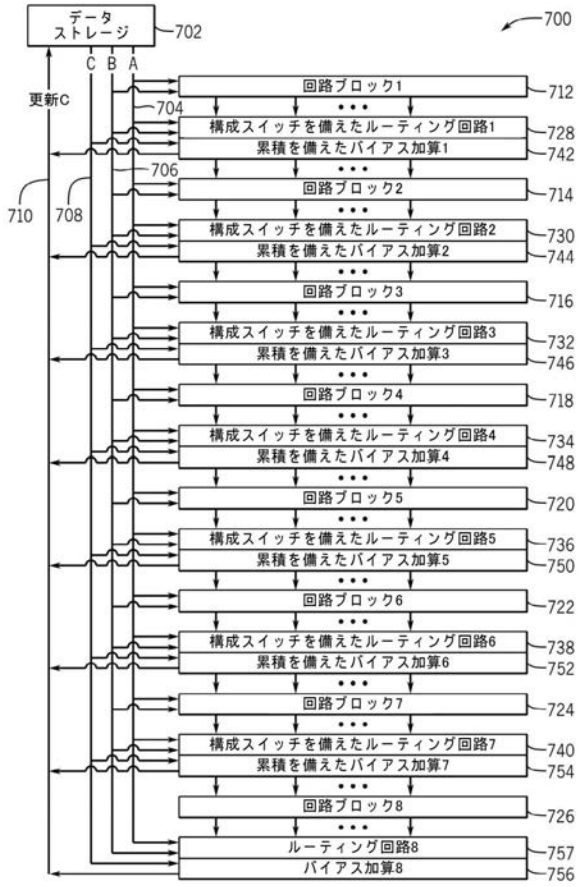
【図５】


行6にデータアウト
PE 行=5
B行列
スイッチがオンの場合、4からのデータインはゼロ
最後のサブタイル
構成スイッチ
構成オン
マルチプレクサ
累積及びバイアス加算
Cタイル
累積イネーブル
バイアス加算
PE 行=4
A行列
B行列
行3からデータイン
累積イネーブル(ISA)
アドレスチェック
Cタイルアドレス

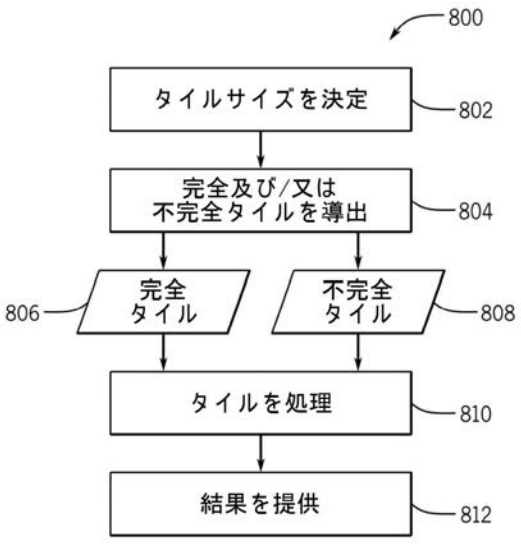
【図６】


バイアス加算
更新C
クロック
バイアス(C)
累積イネーブル
ドット積+バイアス
累積値+バイアス
クロック
最後のサブタイル
アクティベーションイネーブル
カウントリセット
3ビットカウンタ
リセット
累積イネーブル
ドット積

【図７】


データストレージ
C B A
更新C
回路ブロック1
構成スイッチを備えたルーティング回路1
累積を備えたバイアス加算1
回路ブロック2
構成スイッチを備えたルーティング回路2
累積を備えたバイアス加算2
回路ブロック3
構成スイッチを備えたルーティング回路3
累積を備えたバイアス加算3
回路ブロック4
構成スイッチを備えたルーティング回路4
累積を備えたバイアス加算4
回路ブロック5
構成スイッチを備えたルーティング回路5
累積を備えたバイアス加算5
回路ブロック6
構成スイッチを備えたルーティング回路6
累積を備えたバイアス加算6
回路ブロック7
構成スイッチを備えたルーティング回路7
累積を備えたバイアス加算7
回路ブロック8
ルーティング回路8
バイアス加算8

【図８】


タイルサイズを決定
完全及び/又は不完全タイルを導出
完全タイル
不完全タイル
タイルを処理
結果を提供

フロントページの続き

(72)発明者　ガープリート　シン　カルシ
アメリカ合衆国　９５０５４　カリフォルニア州・サンタクララ・ミッション　カレッジ　ブーレバード・２２００　インテル・コーポレーション内
(72)発明者　クリストファー　ジャスティン　ヒューズ
アメリカ合衆国　９５０５４　カリフォルニア州・サンタクララ・ミッション　カレッジ　ブーレバード・２２００　インテル・コーポレーション内

Ｆターム(参考) 5B056 BB42　EE03　FF01　FF02　FF10　FF16

【外国語明細書】
2021108104000001.pdf