(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7561925号
(P7561925)

(45)発行日　令和6年10月4日(2024.10.4)

(24)登録日　令和6年9月26日(2024.9.26)

(51)国際特許分類		FI		
G06F 9/38 (2018.01)		G06F	9/38	370Z
G06F 9/30 (2018.01)		G06F	9/30	350F
G06F 15/173 (2006.01)		G06F	15/173	683B
G06F 15/80 (2006.01)		G06F	15/80	
G06F 17/16 (2006.01)		G06F	17/16	S

請求項の数　20　外国語出願　（全18頁）　最終頁に続く

(21)出願番号　特願2023-114361(P2023-114361)
(22)出願日　令和5年7月12日(2023.7.12)
(62)分割の表示　特願2021-142529(P2021-142529)の分割
原出願日　平成30年5月17日(2018.5.17)
(65)公開番号　特開2023-145517(P2023-145517A)
(43)公開日　令和5年10月11日(2023.10.11)
審査請求日　令和5年8月7日(2023.8.7)
(31)優先権主張番号　62/507,771
(32)優先日　平成29年5月17日(2017.5.17)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　502208397
グーグル エルエルシー
Ｇｏｏｇｌｅ ＬＬＣ
アメリカ合衆国 カリフォルニア州 ９４０４３ マウンテン ビュー アンフィシアター パークウェイ １６００
１６００ Ａｍｐｈｉｔｈｅａｔｒｅ Ｐａｒｋｗａｙ ９４０４３ Ｍｏｕｎｔａｉｎ Ｖｉｅｗ， ＣＡ Ｕ．Ｓ．Ａ．
(74)代理人　110001195
弁理士法人深見特許事務所
(72)発明者　ノリー，トーマス
アメリカ合衆国、９４０４３ カリフォルニア州、マウンテン・ビュー、アンフィシアター・パークウェイ、１６００

最終頁に続く

(54)【発明の名称】　専用ニューラルネットワークトレーニングチップ

(57)【特許請求の範囲】

【請求項１】

専用ハードウェアチップを使用してニューラルネットワークをトレーニングする方法であって、

前記専用ハードウェアチップのベクトルプロセッサが、活性化入力の２次元行列を表すデータを受信することを含み、前記活性化入力の２次元行列は、ニューラルネットワークのそれぞれのネットワーク層に対する入力活性化行列の活性化入力の一部を含み、前記ベクトルプロセッサは、

２次元構造に配置された複数のベクトル処理ユニットと、

前記活性化入力の２次元行列を格納するよう構成されるベクトルレジスタとを含み、前記方法はさらに、

行列乗算ユニットが、前記２次元行列について乗算結果を生成することを含み、前記生成することは、

複数のクロックサイクルの各クロックサイクルにおいて、前記ベクトルプロセッサの前記ベクトルレジスタから前記専用ハードウェアチップの前記行列乗算ユニットに、前記活性化入力の２次元行列における活性化入力のそれぞれの行の少なくとも一部をロードすることと、

前記複数のクロックサイクルの終了前に、前記行列乗算ユニットに重み値をロードすることとを含み、前記重み値は、前記活性化入力の２次元行列における前記活性化入力に対応する形状に配置され、前記生成することはさらに、

前記行列乗算ユニットが、前記活性化入力の２次元行列および重み値を乗算して、前記乗算結果を得ることを含み、前記方法はさらに、
前記乗算結果に基づいて、逆伝播を通じて前記ニューラルネットワークの重み値を更新することを含む、方法。
【請求項２】
前記行列乗算ユニットに、前記活性化入力の２次元行列における前記活性化入力のそれぞれの行の少なくとも前記一部をロードすることは、さらに、前記活性化入力のそれぞれの行の少なくとも前記一部を、第１の浮動小数点フォーマットから、前記第１の浮動小数点フォーマットよりも小さいビットサイズを有する第２の浮動小数点フォーマットに変換することを含む、請求項１に記載の方法。
【請求項３】
前記乗算結果を前記ベクトルレジスタに格納することをさらに含む、請求項１または２に記載の方法。
【請求項４】
前記乗算結果は２次元行列であり、前記乗算結果を格納することは、
別のセットのクロックサイクルのクロックサイクルごとに、前記乗算結果の値のそれぞれの行を、前記行列乗算ユニットにおける先入れ先出しキューに一時的に格納することを含む、請求項１～３のいずれかに記載の方法。
【請求項５】
前記ベクトルプロセッサが、前記専用ハードウェアチップのベクトルメモリから前記活性化入力の２次元行列を表す前記データを受信することをさらに含み、前記ベクトルメモリは、高速のプライベートメモリを前記ベクトルプロセッサに提供するよう構成される、請求項１～４のいずれかに記載の方法。
【請求項６】
前記専用ハードウェアチップの転置ユニットが、前記活性化入力の２次元行列の転置演算を実行することをさらに含む、請求項１～５のいずれかに記載の方法。
【請求項７】
前記専用ハードウェアチップの削減ユニットが、前記乗算結果の数値の削減を実行することと、
前記専用ハードウェアチップの置換ユニットが、前記複数のベクトル処理ユニットに格納された数値を置換することとをさらに含む、請求項１～６のいずれかに記載の方法。
【請求項８】
前記乗算結果を前記専用ハードウェアチップの高帯域幅メモリに格納することをさらに含む、請求項１～７のいずれかに記載の方法。
【請求項９】
前記専用ハードウェアチップの疎計算コアによって、予め構築されたルックアップテーブルを使用して、疎な高次元データを密な低次元データにマッピングすることをさらに含む、請求項１～８のいずれかに記載の方法。
【請求項１０】
前記専用ハードウェアチップのチップ間相互接続を使用して、前記専用ハードウェアチップのインターフェイスまたはリソースを他の専用ハードウェアチップまたはリソースに接続することをさらに含む、請求項１～９のいずれかに記載の方法。
【請求項１１】
前記チップ間相互接続は、前記インターフェイスおよび前記専用ハードウェアチップの高帯域幅メモリを他の専用ハードウェアチップに接続する、請求項１０に記載の方法。
【請求項１２】
前記インターフェイスは、ホストコンピュータへのホストインターフェイス、またはホストコンピュータのネットワークへの標準ネットワークインターフェイスである、請求項１０に記載の方法。
【請求項１３】

前記ベクトルレジスタは３２個のベクトルレジスタを含む、請求項１～１２のいずれかに記載の方法。
【請求項１４】
前記複数のベクトル処理ユニットの各ベクトル処理ユニットは、各クロックサイクルにおいて、２つのそれぞれの算術論理ユニット（ＡＬＵ）命令、それぞれのロード命令、およびそれぞれのストア命令を実行するよう構成される、請求項１～１３のいずれかに記載の方法。
【請求項１５】
前記複数のベクトル処理ユニットにおける各ベクトル処理ユニットは、各クロックサイクルにおいてそれぞれのロード命令およびストア命令を実行するためにそれぞれのオフセットメモリアドレスを演算するよう構成される、請求項１４に記載の方法。
【請求項１６】
前記複数のベクトル処理ユニットの前記２次元構造は、複数のレーンと、前記複数のレーンの各々のための複数のサブレーンとを備え、それぞれのベクトル処理ユニットは、前記複数のサブレーンの各々に位置し、同じレーンに位置する前記複数のベクトル処理ユニットのうちのベクトル処理ユニットは、それぞれのロード命令およびストア命令を介して互いと通信するよう構成される、請求項１～１５のいずれかに記載の方法。
【請求項１７】
１つ以上のコンピュータと１つ以上のストレージデバイスとを備えるシステムであって、前記１つ以上のストレージデバイスには、前記１つ以上のコンピュータによって実行されると前記１つ以上のコンピュータに専用ハードウェアチップを使用してニューラルネットワークをトレーニングするための動作を実行させるよう動作可能である命令が格納され、前記動作は、
前記専用ハードウェアチップのベクトルプロセッサが、活性化入力の２次元行列を表すデータを受信することを含み、前記活性化入力の２次元行列は、ニューラルネットワークのそれぞれのネットワーク層に対する入力活性化行列の活性化入力の一部を含み、前記ベクトルプロセッサは、
２次元構造に配置された複数のベクトル処理ユニットと、
前記活性化入力の２次元行列を格納するよう構成されるベクトルレジスタとを含み、前記動作はさらに、
行列乗算ユニットが、前記２次元行列について乗算結果を生成することを含み、前記生成することは、
複数のクロックサイクルの各クロックサイクルにおいて、前記ベクトルプロセッサの前記ベクトルレジスタから前記専用ハードウェアチップの前記行列乗算ユニットに、前記活性化入力の２次元行列における活性化入力のそれぞれの行の少なくとも一部をロードすることと、
前記複数のクロックサイクルの終了前に、前記行列乗算ユニットに重み値をロードすることとを含み、前記重み値は、前記活性化入力の２次元行列における前記活性化入力に対応する形状に配置され、前記生成することはさらに、
前記行列乗算ユニットが、前記活性化入力の２次元行列および重み値を乗算して、前記乗算結果を得ることを含み、前記動作はさらに、
前記乗算結果に基づいて、逆伝播を通じて前記ニューラルネットワークの重み値を更新することを含む、システム。
【請求項１８】
前記行列乗算ユニットに、前記活性化入力の２次元行列における前記活性化入力のそれぞれの行の少なくとも前記一部をロードすることは、さらに、前記活性化入力のそれぞれの行の少なくとも前記一部を、第１の浮動小数点フォーマットから、第１の浮動小数点フォーマットよりも小さいビットサイズを有する第２の浮動小数点フォーマットに変換することを含む、請求項１７に記載のシステム。
【請求項１９】

１つ以上のコンピュータによって実行されると、前記１つ以上のコンピュータに、専用ハードウェアチップを使用してニューラルネットワークをトレーニングするための動作を実行させる命令を含むプログラムであって、前記動作は、
前記専用ハードウェアチップのベクトルプロセッサが、活性化入力の２次元行列を表すデータを受信することを含み、前記活性化入力の２次元行列は、ニューラルネットワークのそれぞれのネットワーク層に対する入力活性化行列の活性化入力の一部を含み、前記ベクトルプロセッサは、
２次元構造に配置された複数のベクトル処理ユニットと、
前記活性化入力の２次元行列を格納するよう構成されるベクトルレジスタとを含み、前記動作はさらに、
行列乗算ユニットが、前記２次元行列について乗算結果を生成することを含み、前記生成することは、
複数のクロックサイクルの各クロックサイクルにおいて、前記ベクトルプロセッサの前記ベクトルレジスタから前記専用ハードウェアチップの前記行列乗算ユニットに、前記活性化入力の２次元行列における活性化入力のそれぞれの行の少なくとも一部をロードすることと、
前記複数のクロックサイクルの終了前に、前記行列乗算ユニットに重み値をロードすることとを含み、前記重み値は、前記活性化入力の２次元行列における前記活性化入力に対応する形状に配置され、前記生成することはさらに、
前記行列乗算ユニットが、前記活性化入力の２次元行列および重み値を乗算して、前記乗算結果を得ることを含み、前記動作はさらに、
前記乗算結果に基づいて、逆伝播を通じて前記ニューラルネットワークの重み値を更新することを含む、プログラム。
【請求項２０】
前記行列乗算ユニットに、前記活性化入力の２次元行列における前記活性化入力のそれぞれの行の少なくとも前記一部をロードすることは、さらに、前記活性化入力のそれぞれの行の少なくとも前記一部を、第１の浮動小数点フォーマットから、前記第１の浮動小数点フォーマットよりも小さいビットサイズを有する第２の浮動小数点フォーマットに変換することを含む、請求項１９に記載のプログラム。
【発明の詳細な説明】
【背景技術】
【０００１】
背景
この明細書は、ハードウェアでのニューラルネットワーク計算の実行に関する。ニューラルネットワークは機械学習モデルであり、各々が、モデルの１つ以上の層を用いて、受信した入力に対する出力、たとえば分類などを生成する。一部のニューラルネットワークには、出力層に加えて１つ以上の隠れ層が含まれる。各隠れ層の出力は、ネットワーク内の次の層、つまりネットワークの次の隠れ層または出力層への入力として用いられる。ネットワークの各層は、それぞれのパラメータのセットの現在の値に従って、受信した入力から出力を生成する。
【発明の概要】
【０００２】
概要
本明細書では、機械学習ワークロード、特にトレーニング段階に最適化されたプログラム可能な線形代数アクセラレータである専用ハードウェアチップに関する技術について説明する。
【０００３】
一般に、本明細書で説明される主題の革新的な一態様は、特別なハードウェアチップで具現化することができる。
【０００４】

この態様の他の実施形態は、各々、方法のアクションを実行するように構成された、対応するコンピュータシステム、装置、および１つ以上のコンピュータ記憶装置に記録されたコンピュータプログラムを含む。１つ以上のコンピュータのシステムが特定の操作またはアクションを実行するように構成されているとは、動作中にそのシステムにそのような操作またはアクションを実行させるソフトウェア、ファームウェア、ハードウェア、またはそれらの組み合わせがそのシステムにインストールされていることを意味する。１つ以上のコンピュータプログラムが特定の操作またはアクションを実行するように構成されるとは、１つ以上のプログラムが、データ処理装置によって実行されると、データ処理装置にそのような操作またはアクションを実行させる命令を含むことを意味する。

【０００５】

前述の実施形態および他の実施形態は各々、以下の特徴のうちの１つ以上を単独でまたは組み合わせて任意選択で含むことができる。特に、一実施形態は、以下の特徴をすべて組み合わせて含む。

【０００６】

ニューラルネットワークをトレーニングするための専用ハードウェアチップは、専用ハードウェアチップの計算動作を制御するように構成されたスカラープロセッサと、ベクトル処理ユニットの２次元配列を有するように構成されたベクトルプロセッサとを備えてもよく、ベクトル処理ユニットは、すべて、同じ命令を単一命令複数データ方式で実行し、ベクトルプロセッサのロードおよびストア命令を通して互いに通信し、専用ハードウェアチップはさらに、ベクトルプロセッサに結合され、乗算結果を得るために、少なくとも１つの２次元行列を第２の１次元ベクトルまたは２次元行列と乗算するように構成された行列乗算ユニットを備えてもよい。

【０００７】

ベクトルメモリが、ベクトルプロセッサに高速のプライベートメモリを提供するように構成されてもよい。スカラーメモリが、スカラープロセッサに高速のプライベートメモリを提供するように構成されてもよい。転置ユニットが、行列の転置演算を実行するように構成されてもよい。削減および置換ユニットが、ベクトルアレイの異なるレーン間において、数値上で削減を実行し、数値を置換するように構成されてもよい。高帯域幅メモリが、専用ハードウェアチップのデータを記憶するように構成されてもよい。専用ハードウェアチップは、疎計算コアを備えてもよい。

【０００８】

専用ハードウェアチップは、インターフェイスと、専用ハードウェアチップ上のインターフェイスまたはリソースを他の専用ハードウェアチップまたはリソースに接続するチップ間相互接続とを備えてもよい。

【０００９】

専用ハードウェアチップは、高帯域幅メモリを備えてもよい。チップ間相互接続は、インターフェイスおよび高帯域幅メモリを他の専用ハードウェアチップに接続してもよい。インターフェイスは、ホストコンピュータへのホストインターフェイスであってもよい。インターフェイスは、ホストコンピュータのネットワークへの標準ネットワークインターフェイスであってもよい。

【００１０】

この明細書において記載される主題は、以下の利点の１つ以上を実現するように特定の実施の形態において実現することができる。専用ハードウェアチップは、機械学習用の３２ビット以下の精度計算用に最適化されながらも、従来の０次元および１次元のテンソル計算に加えて、より高次元のテンソル（つまり、２次元以上）をネイティブにサポートするプロセッサを含む。

【００１１】

この明細書の主題の１つ以上の実施形態の詳細は、添付の図面および以下の詳細な説明において述べられる。主題の他の特徴、局面および利点は、詳細な説明、図面および特許請求の範囲から明らかになる。

【図面の簡単な説明】
【００１２】
【図１】ボード上において円形トポロジで接続された専用ハードウェアチップの例示的集合体を接続する高速接続の例示的トポロジを示す。
【図２】ニューラルネットワークをトレーニングするための例示的な専用ハードウェアチップのハイレベル図を示す。
【図３】コンピュートコアのハイレベルの例を示す。
【図４】ニューラルネットワークのためにトレーニングを実行するチップのより詳細な図を示す。
【発明を実施するための形態】
【００１３】
さまざまな図面の同様の参照番号および名称は、同様の要素を示す。
詳細な説明
複数の層を有するニューラルネットワークをトレーニングし、それを推論の計算に用いることができる。一般に、ニューラルネットワークの一部またはすべての層は、ニューラルネットワークのトレーニング中に調整されるパラメータを有する。たとえば、一部またはすべての層は、その層に対する、重みとも称されるパラメータの行列に、層出力の生成の一部として、その層への入力を乗算できる。行列内のパラメータの値は、ニューラルネットワークのトレーニング中に調整される。
【００１４】
特に、トレーニング中、トレーニングシステムは、ニューラルネットワークのトレーニング手順を実行して、ニューラルネットワークのパラメータの値を調整して、たとえば、パラメータの初期値からパラメータのトレーニングを経た値を判断する。トレーニングシステムは、逆伝播として知られる誤差の逆伝播を、最適化方法と組み合わせて用いて、ニューラルネットワークの各パラメータに関して目的関数の勾配を計算し、その勾配を用いてパラメータの値を調整する。
【００１５】
トレーニングされたニューラルネットワークは、順方向伝播を用いて推論を計算でき、つまり、ニューラルネットワークの層を介して入力を処理し、その入力に対するニューラルネットワーク出力を生成できる。
【００１６】
たとえば、入力が与えられると、ニューラルネットワークはその入力に対する推論を計算することができる。ニューラルネットワークは、ニューラルネットワークの各層を通して入力を処理することによって、この推論を計算する。いくつかの実装形態では、ニューラルネットワークの層はシーケンスで配置される。
【００１７】
したがって、受け取った入力から推論を計算するために、ニューラルネットワークはその入力を受け取り、それを各ニューラルネットワーク層を通してシーケンスで処理して推論を生成し、１つのニューラルネットワーク層からの出力が次のニューラルネットワーク層への入力として与えられる。ニューラルネットワーク層へのデータ入力、たとえば、ニューラルネットワークへの入力、またはシーケンス内におけるその層の下の層の、あるニューラルネットワーク層への出力は、その層への活性化入力と呼ぶことができる。
【００１８】
いくつかの実装形態では、ニューラルネットワークの層は有向グラフで配置される。つまり、任意の特定の層が複数の入力、複数の出力、またはそれらの両方を受け取ることができる。ニューラルネットワークの層は、ある層の出力を前の層への入力として送り返すことができるように構成することもできる。
【００１９】
ある例示的システムは、行列乗算やその他の多次元配列の計算に最適化された高性能マルチチップテンソル計算システムである。これらの操作は、ニューラルネットワークのト

レーニング、およびオプションで、ニューラルネットワークを用いて推論を計算するために重要である。

【００２０】

ある例示的システムでは、システムがトレーニングおよび推論計算を効率的に実行するために、複数の専用チップが操作を分散するように配置される。１つの実装形態では、ボード上に４つのチップがあり、より大きなシステムでは、多くのボードがラック内で隣り合っているか、またはそうでなければ相互にデータ通信を行なう。

【００２１】

図１は、ボード上において円形トポロジで接続された専用ハードウェアチップ１０１ａ～１０１ｄの例示的集合体を接続する高速接続の例示的トポロジを示す。各チップには２つのプロセッサ（１０２ａ～１０２ｈ）が含まれる。このトポロジは、１次元（１Ｄ）トーラスであり、１Ｄトーラスでは、各チップは２つの隣接チップに直接接続される。示されているように、一部の実装形態では、チップには、操作を実行するようソフトウェア命令またはファームウェア命令でプログラムされたマイクロプロセッサコアが含まれる。図１では、すべてのチップが単一のモジュール１００上にある。図に示されているプロセッサ間の線は、高速データ通信リンクを表す。プロセッサは１つの集積回路基板上に有利に製造されるが、複数の基板上に製造することもできる。チップ境界を越えて、リンクは、チップ間ネットワークリンクであり、同じチップ上のプロセッサは、チップ内インターフェイスリンクを介して通信する。リンクは、一度に１つのプロセッサのみがデータを送信できる半二重リンクであってもよいし、データを両方向に同時に送信できる全二重リンクであってもよい。この例示的トポロジを用いる並列処理およびその他については、２０１７年２月２１日に提出され、ここに引用により援用される、「PARALLEL PROCESSING OF REDUCTION AND BROADCAST OPERATIONS ON LARGE DATASETS OF NON-SCALAR DATA（非スカラーデータの大規模データセットの削減およびブロードキャスト操作の並列処理）」と題される米国特許出願第６２／４６１，７５８号において詳細に説明される。

【００２２】

図２は、ニューラルネットワークをトレーニングするための例示的な専用ハードウェアチップのハイレベル図を示す。図示されているように、単一の専用ハードウェアチップには２つの独立したプロセッサ（２０２ａ、２０２ｂ）が含まれる。各プロセッサ（２０２ａ、２０２ｂ）には、２つの異なるコア：（１）コンピュートコア、たとえば非常に長い命令語（ＶＬＩＷ）マシン（２０３ａ、２０３ｂ）、および（２）疎計算コア、つまり埋め込み層アクセラレータ（２０５ａ、２０５ｂ）が含まれる。

【００２３】

各コア（２０３ａ、２０３ｂ）は、密線形代数問題に対して最適化される。単一の非常に長い命令語が、いくつかのコンピュートコアを並列で制御する。コンピュートコアは、図３および図４を参照してより詳細に説明される。

【００２４】

例示的な疎計算コア（２０５ａ、２０５ｂ）は、非常に疎である高次元データを密な低次元データにマッピングし、残りの層が密に充填された入力データを処理するようにする。たとえば、疎計算コアは、トレーニング中のニューラルネットワークの埋め込み層の計算を実行できる。

【００２５】

この疎から密へのマッピングを実行するために、疎計算コアは、埋め込みテーブルである、予め構築されたルックアップテーブルを用いる。たとえば、ユーザ入力として一連のクエリワードがある場合、各クエリワードはハッシュ識別子またはワンホットエンコードされたベクトルに変換される。識別子をテーブルインデックスとして用いて、埋め込みテーブルは対応する密なベクトルを返し、それは、次の層への入力活性化ベクトルになることができる。疎計算コアは、検索クエリワードにわたって削減操作を実行して、１つの密な活性化ベクトルを作成することもできる。疎計算コアは、効率的な疎の、分散されたル

ックアップを実行し、なぜならば、埋め込みテーブルが巨大であり得、専用ハードウェアチップの１つの限られた容量の高帯域幅メモリに収まらないためである。疎計算コア機能に関する詳細は、２０１６年２月５日に提出された「MATRIX PROCESSING APPARATUS（行
列処理装置）」と題される米国特許出願第１５／０１６，４８６号に記載され、それをここに引用により援用する。

【００２６】

図３は、コンピュートコア（３００）のハイレベルの例を示す。コンピュートコアは、複数のコンピュートユニットを並列で制御するマシン、つまりＶＬＩＷマシンであることができる。各コンピュートコア（３００）には、スカラーメモリ（３０４）、ベクトルメモリ（３０８）、スカラープロセッサ（３０３）、ベクトルプロセッサ（３０６）、および拡張ベクトルユニット（つまり、行列乗算ユニット（ＭＸＵ）（３１３）と、転置ユニット（ＸＵ）（３１４）と、削減および置換ユニット（ＲＰＵ）（３１６））が含まれる。

【００２７】

例示的なスカラープロセッサは、ＶＬＩＷ命令のフェッチ／実行ループを実行し、コンピュートコアを制御する。命令バンドルをフェッチしてデコードした後、スカラープロセッサ自体は、スカラープロセッサ（３０３）およびスカラーメモリ（３０４）の複数のマルチビットレジスタ、つまり３２の３２ビットレジスタを用いて、命令バンドルのスカラースロットにおいて見いだされる命令を実行するのみである。スカラー命令セットには、アドレス計算、ロード／ストア命令、分岐命令などで使用される通常の算術演算が含まれる。残りの命令スロットは、ベクトルプロセッサ（３０６）または他の拡張ベクトルユニット（３１３、３１４、３１６）の命令をエンコードする。デコードされたベクトル命令は、ベクトルプロセッサ（３０６）に転送される。

【００２８】

ベクトル命令とともに、スカラープロセッサ（３０３）は、最大３つのスカラーレジスタの値を他のプロセッサおよびユニットに転送して、演算を実行できる。スカラープロセッサは、ベクトルプロセッサから計算結果を直接取得することもできる。ただし、一部の実装形態では、例示的なチップは、ベクトルプロセッサからスカラープロセッサへの低帯域幅通信パスを有する。

【００２９】

ベクトル命令ディスパッチャは、スカラープロセッサとベクトルプロセッサとの間にある。このディスパッチャは、非スカラーＶＬＩＷスロットからデコードされた命令を受け取り、それらの命令をベクトルプロセッサ（３０６）にブロードキャストする。ベクトルプロセッサ（３０６）は、同じ命令を単一命令複数データ（ＳＩＭＤ）方式で実行するベクトル処理ユニットの２次元配列、すなわち１２８×８の配列からなる。ベクトル処理ユニットは、図４を参照して詳細に説明される。

【００３０】

例示的なスカラープロセッサ（３０３）は、小さい、高速な、プライベートスカラーメモリ（３０４）にアクセスする。このメモリは、はるかにより大きく、低速な高帯域幅メモリ（ＨＢＭ）（３１０）によってバックアップされる。同様に、例示的なベクトルプロセッサ（３０６）は、小さい、高速な、プライベートベクトルメモリ（３０６）にアクセスする。このメモリも、ＨＢＭ（３１０）によってバックアップされる。ワード粒度アクセスは、スカラープロセッサ（３０３）とスカラーメモリ（３０４）との間、またはベクトルプロセッサ（３０６）とベクトルメモリ（３０８）との間で起こる。ベクトルプロセッサとベクトルメモリとの間のロードおよびストアの粒度は、１２８個の３２ビットワードのベクトルである。ダイレクトメモリアクセスは、スカラーメモリ（３０４）とＨＢＭ（３１０）との間、およびベクトルメモリ（３０６）とＨＢＭ（３１０）との間で起こる。いくつかの実装形態では、ＨＢＭ（３１０）からプロセッサ（３０３、３０６）へのメモリ転送は、スカラーメモリまたはベクトルメモリを介してのみ実行できる。さらに、スカラーメモリとベクトルメモリとの間でダイレクトメモリ転送が行われない場合がある。

【００３１】
命令は、拡張ベクトルユニット演算を指定する場合がある。実行された各ベクトルユニット命令に加えて、各々が１つのレジスタ値を拡張ベクトルユニットに入力オペランドとして送ることができる２次元、つまり１２８×８のベクトルユニットがある。各拡張ベクトルユニットは、入力オペランドを受け取り、対応する演算を実行し、結果をベクトルプロセッサ（３０６）に返す。拡張ベクトルユニットは、図４を参照して以下に説明される。
【００３２】
図４は、ニューラルネットワークのためにトレーニングを実行するチップのより詳細な図を示す。上で図示し説明したように、チップには２つのコンピュートコア（４８０ａ、４８０ｂ）と２つの疎計算コア（４５２ａ、４５２ｂ）とが含まれる。
【００３３】
チップには、ホストコンピュータ（４５０）または複数のホストコンピュータへのインターフェイスを含む共有領域がある。このインターフェイスは、ホストコンピュータへのホストインターフェイス、またはホストコンピュータのネットワークへの標準ネットワークインターフェイスとすることができる。共有領域は、下部に沿って高帯域幅メモリのスタック（４５６ａ～４５６ｄ）、およびインターフェイスとメモリとを接続するチップ間相互接続（４４８）、ならびに他のチップからのデータも有し得る。相互接続は、インターフェイスをハードウェアチップ上の計算リソースに接続することもできる。高帯域幅メモリの複数のスタック、つまり２つのスタック（４５６ａ～４５６ｂ、４５６ｃ～４５６ｄ）が各コンピュートコア（４８０ａ、４８０ｂ）に関連付けられる。
【００３４】
チップは、データを高帯域幅メモリ（４５６ｃ～４５６ｄ）に保存し、そのデータをベクトルメモリにおいて読込および読出し（４４６）、そのデータを処理する。コンピュートコア（４８０ｂ）自体は、２次元に分割されたオンチップＳ－ＲＡＭであるベクトルメモリ（４４６）を含む。ベクトルメモリには、アドレスが浮動小数点数、つまり各々３２ビットである１２８個の数値を保持するアドレス空間がある。コンピュートコア（４８０ｂ）は、値を計算する計算ユニット、および計算ユニットを制御するスカラーユニットも含む。計算ユニットはベクトルプロセッサを含んでもよく、スカラーユニットはスカラープロセッサを含んでもよい。専用チップの一部を形成し得るコンピュートコアは、行列乗算ユニット、または行列、つまり１２８×１２８の行列の転置演算を実行する転置ユニット（４２２）などの別の拡張演算ユニット、ならびに削減および置換ユニットをさらに含むことができる。
【００３５】
ベクトルプロセッサ（３０６）は、ベクトル処理ユニットの２次元配列、すなわち１２８×８から成り、これらはすべて、同じ命令を単一命令複数データ（ＳＩＭＤ）方式で実行する。ベクトルプロセッサには、レーンとサブレーン、つまり１２８本のレーンおよび８本のサブレーンがある。レーン内において、ベクトルユニットはロード命令およびストア命令を介して互いに通信する。各ベクトルユニットは、一度に１つの４バイト値にアクセスできる。同じレーンに属さないベクトルユニットは直接通信できない。これらのベクトルユニットは、以下で説明する削減／置換ユニットを用いる必要がある。
【００３６】
計算ユニットは、ベクトル処理ユニットにおいて、浮動小数点演算および整数演算の両方に用いることができるベクトルレジスタ（４４０）、つまり３２個のレジスタを含む。計算ユニットは、計算を実行するための２つの算術論理ユニット（ＡＬＵ）（４０６ｃ～４０６ｄ）を含む。一方のＡＬＵ（４０６ｃ）は浮動小数点加算を実行し、他方のＡＬＵ（４０６ｄ）は浮動小数点乗算を実行する。両方のＡＬＵ（４０６ｃ～４０６ｄ）は、シフト、マスク、比較などの他のさまざまな演算を実行できる。たとえば、コンピュートコア（４８０ｂ）は、ベクトルレジスタＶ１と第２のベクトルレジスタＶ２とを加算し、結果を第３のベクトルレジスタＶ３に入れたい場合がある。この加算を計算するために、コンピュートコア（４８０ｂ）は複数の演算を１クロックサイクルで実行する。これらのレ

ジスタをオペランドとして用いて、各ベクトルユニットは、クロックサイクルごとに２つのＡＬＵ命令と１つのロードおよび１つのストア命令とを同時に実行できる。ロードまたはストア命令のベースアドレスは、スカラープロセッサで計算でき、ベクトルプロセッサに転送できる。各サブレーンにおけるベクトルユニットの各々は、ストライドや特別なインデックス付きアドレスレジスタなどのさまざまな方法を用いて、それ自体のオフセットアドレスを計算できる。

【００３７】

計算ユニットは、平方根や逆数などの演算を実行する拡張単項パイプライン（ＥＵＰ）（４１６）も含む。コンピュートコア（４８０ｂ）は、これらの演算を実行するのに３クロックサイクルかかり、なぜならば、それらは計算が複雑であるからである。ＥＵＰ処理には１クロックサイクル以上かかるため、結果を保存する先入れ先出しのデータストレージがある。演算が終了すると、結果はＦＩＦＯに保存される。コンピュートコアは、後で別の命令を用いて、ＦＩＦＯからデータを引き出し、それをベクトルレジスタに格納できる。乱数生成器（４２０）により、コンピュートコア（４８０ｂ）はサイクルごとに複数の乱数、つまりサイクルごとに１２８の乱数を生成できる。

【００３８】

上記のように、専用のハードウェアチップの一部として実装できる各プロセッサは、３つの拡張演算ユニット、つまり、行列乗算演算を実行する行列乗算ユニット（４４８）と、行列、すなわち１２８×１２８の行列の転置演算を実行する転置ユニット（４２２）と、削減および置換ユニット（図４において別個のユニット４２４、４２６として示される）とを有する。

【００３９】

行列乗算ユニットは、２つの行列間で行列乗算を実行する。コンピュートコアは、乗算される行列である一連の数値を読み込む必要があるため、行列乗算ユニット（４３８）はデータを取り込む。図示されているように、データはベクトルレジスタ（４４０）から来る。各ベクトルレジスタには、１２８×８の数値、つまり３２ビットの数値が含まれる。しかしながら、データを行列乗算ユニット（４４８）に送って、数値をより小さなビットサイズ、つまり３２ビットから１６ビットに変更すると、浮動小数点変換が発生する場合がある。並直列変換器（４４０）は、ベクトルレジスタから数値が読み取られるときに、２次元配列つまり１２８×８の行列が１２８個の数値のセットとして読み取られ、次の８クロックサイクルの各々ごとに行列乗算ユニット（４４８）に送信されることを保証する。行列乗算がその計算を完了した後、結果は非直列化され（４４２ａ、４４２ｂ）、これは、結果行列が、ある数のクロックサイクルの間保持されることを意味する。たとえば、１２８×８の配列の場合、１２８個の数値が８クロックサイクルの各々ごとに保持され、次いでＦＩＦＯにプッシュされ、１２８×８の数値の２次元配列を１クロックサイクルで取得してベクトルレジスタ（４４０）に格納できる。

【００４０】

複数すなわち１２８のサイクルからなる期間にわたって、重みが、行列を乗算する数値として行列乗算ユニット（４４８）にシフトされる。行列および重みがロードされると、コンピュートコア（４８０）は、数値のセット、つまり１２８×８を行列乗算ユニット（４４８）に送ることができる。セットの各ラインに行列を乗算して、クロックサイクルごとにある数の結果、つまり１２８を生成できる。コンピュートコアが行列乗算を実行している間、コンピュートコアは、前の行列の計算プロセスが完了したときに、コンピュートコアが乗算する次の行列が利用できるように、バックグラウンドにおいて次の行列になる新たな数値のセットのシフトも行う。行列乗算ユニット（４４８）は、「LOW MATRIX MULTIPLY UNIT COMPOSED OF MULTI-BIT CELLS（マルチビットセルで構成された低行列乗算ユニット）」と題される１６１１３－８２５１００１、および「MATRIX MULTIPLY UNIT WITH NUMERICS OPTIMIZED FOR NEURAL NETWORK APPLICATIONS（数値がニューラルネットワークアプリケーション向けに最適化された行列乗算ユニット）」と題される１６１１３－８２５２００１に、より詳細に説明され、それらの両方

をここに引用により援用する。

【００４１】

転置ユニットは、行列を転置する。転置ユニット（４２２）は、数値を取り込み、それらを転置して、レーンを横切る数が他の次元の数値と転置されるようにする。一部の実装形態では、ベクトルプロセッサは１２８×８のベクトルユニットを含む。したがって、１２８×１２８の行列を転置するには、完全な行列転置のために１６個の個別の転置命令が必要である。転置が終了すると、転置された行列が利用可能になる。ただし、転置された行列をベクトルレジスタファイルに移動するには、明示的な命令が必要である。

【００４２】

削減／置換ユニット（またはユニット４２４、４２６）は、置換、レーン回転、回転置換、レーン削減、置換されたレーン削減、およびセグメント化された置換されたレーン削減などのさまざまな操作をサポートすることで、クロスレーン通信の問題に対処する。図示されているように、これらの計算は別々であるが、コンピュートコアは一方もしくは他方または一方に連鎖された他方を用いることができる。削減ユニット（４２４）は、数値からなる各ラインにおけるすべての数値を加算し、それらの数値を置換ユニット（４２６）に供給する。置換ユニットは、異なるレーン間でデータを移動する。転置ユニット、削減ユニット、置換ユニット、および行列乗算ユニットは、各々、完了までに１クロックサイクル以上かかる。したがって、各ユニットにはＦＩＦＯが関連付けられ、計算結果をＦＩＦＯにプッシュし、後で別の命令を実行して、データをＦＩＦＯからベクトルレジスタにプルできる。ＦＩＦＯを用いることにより、コンピュートコアは、長々とした演算の間、複数のベクトルレジスタを予約する必要がない。図示されているように、各ユニットはベクトルレジスタ（４４０）からデータを取得する。

【００４３】

コンピュートコアは、スカラーユニットを用いて計算ユニットを制御する。スカラーユニットには２つの主要な機能があり、それは、（１）ループカウントおよびアドレス指定を実行すること、ならびに（２）ＤＭＡコントローラがバックグラウンドにおいて高帯域幅メモリ（４５６ｃ～４５６ｄ）とベクトルメモリ（４４６）との間で、および次いで例示のシステムにおける他のチップへのチップ間接続（４４８）にデータを移動するよう、ダイレクトメモリアドレス（ＤＭＡ）要求を生成することである。スカラーユニットは、命令メモリ（４０４）と、命令デコードおよび発行（４０２）と、スカラーレジスタすなわち３２ビットを含むスカラー処理ユニット（４０８）と、スカラーメモリ（４１０）と、クロックサイクルごとに２つの演算を実行する２つのＡＬＵ（４０６ａ、４０６ｂ）とを含む。スカラーユニットは、オペランドと即値とをベクトル演算に渡すことができる。各命令は、ベクトルレジスタ（４４０）で実行される命令を含む命令バンドルとして、命令デコードおよび発行（４０２）から送ることができる。各命令バンドルは非常に長い命令語（ＶＬＩＷ）であり、各命令はある数のビット幅であり、ある数の命令フィールドに分割される。

【００４４】

チップ４００を用いて、ニューラルネットワークのトレーニングの少なくとも一部を実行することができる。特に、ニューラルネットワークをトレーニングする場合、システムはホストインターフェイス（４５０）を用いてホストコンピュータからラベル付きトレーニングデータを受信する。ホストインターフェイスは、ニューラルネットワーク計算のためのパラメータを含む命令を受信することもできる。パラメータは、処理すべき層の数、各層についての対応する重み入力のセット、活性化入力の初期セット、つまり推論の計算またはトレーニングの対象となるニューラルネットワークへの入力であるトレーニングデータ、各層の対応する入力および出力サイズ、ニューラルネットワーク計算のストライド値、ならびに処理対象の層のタイプ、たとえば畳み込み層または全結合層、のうちの少なくとも１つ以上を含むことができる。

【００４５】

重み入力のセットおよび活性化入力のセットは、コンピュートコアの行列乗算ユニット

に送ることができる。重み入力および活性化入力を行列乗算ユニットに送る前に、システム内の他のコンポーネントが入力に対して他の計算を実行してもよい。一部の実装形態では、疎計算コアからコンピュートコアに活性化を送る方法が２つある。まず、疎計算コアは、高帯域幅メモリを介して通信を送信することができる。大量のデータの場合、疎計算コアは、ダイレクトメモリアドレス（ＤＭＡ）命令を用いて活性化を高帯域幅メモリに格納でき、これにより、コンピュートコアにおいてターゲット同期フラグが更新される。コンピュートコアは、同期命令を用いてこの同期フラグを待つことができる。同期フラグがセットされると、計算コアはＤＭＡ命令を用いて、活性化を高帯域幅メモリから対応するベクトルメモリにコピーする。

【００４６】

次に、疎計算コアは、通信をコンピュートコアベクトルメモリに直接送信できる。データ量が大きくない場合（つまり、コンピュートコアベクトルメモリに収まる場合）、疎計算コアは、コンピュートコアに同期フラグで通知しながら、ＤＭＡ命令を用いてコンピュートコアのベクトルメモリに活性化を直接格納できる。コンピュートコアは、この同期フラグを待ったのち、活性化に依存する計算を実行することができる。

【００４７】

行列乗算ユニットは、重み入力および活性化入力を処理し、出力のベクトルまたは行列をベクトル処理ユニットに与えることができる。ベクトル処理ユニットは、処理された出力のベクトルまたは行列を格納できる。たとえば、ベクトル処理ユニットは、非線形関数を行列乗算ユニットの出力に適用して、活性化された値を生成できる。いくつかの実装形態では、ベクトル処理ユニットは、正規化された値、プールされた値、またはその両方を生成する。処理された出力のベクトルは、たとえばニューラルネットワーク内の後続の層で用いるために、行列乗算ユニットへの活性化入力として用いることができる。

【００４８】

トレーニングデータのバッチについての処理済み出力のベクトルが計算されると、それらの出力をラベル付きトレーニングデータの期待される出力と比較して、誤差を判断できる。その後、システムは、ネットワークをトレーニングするために、逆伝播を実行して、ニューラルネットワークを介して誤差を伝播できる。損失関数の勾配は、オンチップでベクトル処理ユニットの算術論理ユニットを用いて計算される。

【００４９】

ある例示的システムでは、ニューラルネットワークを介した逆伝播を実行するために、活性化勾配が必要である。活性化勾配をコンピュートコアから疎計算コアに送るために、例示のシステムでは、コンピュートコアＤＭＡ命令を用いて、ターゲット疎計算コアに同期フラグで通知しながら、活性化勾配を高帯域幅メモリに保存できる。疎計算コアは、この同期フラグを待ったのち、活性化勾配に依存する計算を実行することができる。

【００５０】

行列乗算ユニットは、逆伝播のために２つの行列乗算演算を実行する。一方の行列乗算は、逆伝播誤差をネットワーク内の前の層からネットワークを通る逆方向パスに沿って重みに適用して、重みを調整してニューラルネットワークのための新たな重みを決定する。第２の行列乗算は、ニューラルネットワーク内の前の層へのフィードバックとして、元の活性化に誤差を適用する。元の活性化は、順方向パス中に生成され、逆方向パス中に用いるために保存されてもよい。計算には、浮動小数点加算、減算、および乗算を含む、ベクトル処理ユニットにおける汎用命令を用いることができる。汎用命令には、比較、シフト、マスク、および論理演算も含まれ得る。行列の乗算は非常に加速され得るが、ベクトル処理ユニットの算術論理ユニットは、サイクルあたり、コアあたり１２８×８×２の演算の速度で一般的な計算を実行する。

【００５１】

本明細書において記載される主題および機能的動作の実施形態は、本明細書に開示される構造およびそれらの構造的等価物を含む、デジタル電子回路系において、有形で実施されるコンピュータソフトウェアもしくはファームウェアにおいて、コンピュータハードウ

ェアにおいて、またはそれらの１つ以上の組合せにおいて実現され得る。本明細書に記載される主題の実施形態は、１つ以上のコンピュータプログラムとして、すなわち、データ処理装置による実行のために、または、データ処理装置の動作を制御するために有形の非一時的な記憶媒体上でエンコードされたコンピュータプログラム命令の１つ以上のモジュールとして実現され得る。コンピュータ記憶媒体は、機械可読記憶装置、機械可読記憶基板、ランダムもしくはシリアルアクセスメモリデバイス、または、それらの１つ以上の組合せであり得る。代替的に、または加えて、プログラム命令は、データ処理装置による実行に対して好適な受信側装置への送信のために情報をエンコードするよう生成される、たとえばマシンにより生成された電気信号、光信号、または電磁気信号などの、人為的に生成された伝搬される信号上でエンコードすることができる。

【００５２】

「データ処理装置」という用語は、データ処理ハードウェアを指し、例としてプログラマブルプロセッサ、コンピュータ、または複数のプロセッサもしくはコンピュータを含む、データを処理するためのすべての種類の装置、デバイスおよびマシンを包含する。当該装置は、たとえばＦＰＧＡ（フィールドプログラマブルゲートアレイ）またはＡＳＩＣ（特定用途向け集積回路）といった特定目的論理回路でもあることができるかまたはそれをさらに含み得る。当該装置は、ハードウェアに加えて、たとえばプロセッサファームウェア、プロトコルスタック、データベース管理システム、オペレーティングシステム、または、それらの１つ以上の組合せを構成するコードといった、コンピュータプログラムについて実行環境を作成するコードをオプションとして含み得る。

【００５３】

プログラム、ソフトウェア、ソフトウェアアプリケーション、アプリケーション、モジュール、ソフトウェアモジュール、スクリプトまたはコードとも称され、または記載され得るコンピュータプログラムは、コンパイル型もしくはインタープリタ型言語、または宣言型もしくは手続き型言語を含む任意の形態のプログラミング言語で記述され得、スタンドアロンプログラムとして、または、モジュール、コンポーネント、サブルーチン、もしくは、コンピューティング環境で使用するのに好適な他のユニットとして任意の形態で展開され得る。プログラムは、ファイルシステムにおけるファイルに対応し得るが、対応する必要があるわけではない。プログラムは、当該プログラムに専用である単一のファイルにおいて、または、複数の連携ファイル（ｃｏｏｒｄｉｎａｔｅｄ　ｆｉｌｅｓ）（たとえばコードの１つ以上のモジュール、サブプログラムまたは部分を格納するファイル）において、他のプログラムまたはデータ（たとえばマークアップ言語ドキュメントに格納される１つ以上のスクリプト）を保持するファイルの一部に格納され得る。コンピュータプログラムは、１つの場所に位置するかもしくは複数の場所にわたって分散されデータ通信ネットワークによって相互接続される１つのコンピュータまたは複数のコンピュータ上で実行されるように展開され得る。

【００５４】

本明細書に記載されるプロセスおよび論理フローは、入力データ上で動作し出力を生成することにより機能を実行するよう１つ以上のプログラマブルコンピュータが１つ以上のコンピュータプログラムを実行することによって実行され得る。本プロセスおよび論理フローは、たとえばＦＰＧＡもしくはＡＳＩＣといった特殊目的論理回路系によっても、または特殊目的論理回路計と１つ以上のプログラムされたコンピュータとの組み合わせによっても実行され得る。

【００５５】

コンピュータプログラムの実行に好適であるコンピュータは、汎用マイクロプロセッサもしくは特殊目的マイクロプロセッサもしくはその両方または任意の種類の中央処理ユニットに基づき得る。一般に、中央処理ユニットは、リードオンリメモリもしくはランダムアクセスメモリまたはその両方から命令およびデータを受け取る。コンピュータの必須の要素は、命令を実行するための中央処理ユニットと、命令およびデータを格納するための１つ以上のメモリデバイスとである。中央処理ユニットおよびメモリは、特殊目的論理回

路系によって補足され得るか、または特殊目的論理回路に組み込まれ得る。一般に、コンピュータはさらに、たとえば磁気ディスク、光磁気ディスクまたは光ディスクといった、データを格納するための１つ以上の大容量記憶装置を含むか、当該１つ以上の大容量記憶装置からデータを受け取るかもしくは当該１つ以上の大容量記憶装置にデータを転送するよう動作可能に結合されるか、またはその両方を行う。しかしながら、コンピュータはそのような装置を有する必要はない。さらに、コンピュータはたとえば、携帯電話、携帯情報端末（ＰＤＡ）、モバイルオーディオまたはビデオプレーヤ、ゲームコンソール、全地球測位システム（ＧＰＳ）受信機、またはポータブル記憶装置（たとえばユニバーサルシリアルバス（ＵＳＢ）フラッシュドライブ）といった別のデバイスに埋め込まれ得る。

【００５６】

コンピュータプログラム命令およびデータを格納するのに好適であるコンピュータ可読媒体は、例として、たとえばＥＰＲＯＭ、ＥＥＰＲＯＭおよびフラッシュメモリデバイスといった半導体メモリデバイスを含むすべての形態の不揮発性メモリ、媒体およびメモリデバイス；たとえば内部ハードディスクまたはリムーバブルディスクといった磁気ディスク；光磁気ディスク；ならびにＣＤ－ＲＯＭおよびＤＶＤ－ＲＯＭディスクを含む。

【００５７】

ユーザとの対話を与えるために、本明細書に記載される主題の実施形態は、たとえばＣＲＴ（陰極線管）またはＬＣＤ（液晶ディスプレイ）モニタといったユーザに対して情報を表示するための表示デバイスと、たとえばマウス、トラックボールといったユーザがコンピュータに入力を与えることができるキーボードおよびポインティングデバイスとを有するコンピュータ上で実現され得る。他の種類のデバイスが、同様に、ユーザとの対話を与えるために用いられ得；たとえば、ユーザに提供されるフィードバックは、たとえば視覚フィードバック、聴覚フィードバックまたは触覚フィードバックといった任意の形態の感覚フィードバックであり得；ユーザからの入力は、音響入力、音声入力、または触覚入力を含む任意の形態で受け取られ得る。加えて、コンピュータは、ユーザが使用するデバイスにドキュメントを送信しユーザが使用するデバイスからドキュメントを受信することによって、たとえば、ウェブブラウザから受信された要求に応答してユーザのデバイス上のウェブブラウザにウェブページを送信することによって、ユーザと対話し得る。また、コンピュータは、テキストメッセージまたは他の形式のメッセージを個人用デバイス、たとえばスマートフォンなどに送信し、メッセージングアプリケーションを実行し、ユーザから応答メッセージを受信することにより、ユーザと対話できる。

【００５８】

本明細書に記載される主題の実施形態は、たとえばデータサーバとしてバックエンドコンポーネントを含む計算システムにおいて実現され得るか、たとえばアプリケーションサーバといったミドルウェアコンポーネントを含む計算システムにおいて実現され得るか、たとえば本明細書に記載される主題の実現例とユーザが対話することが可能であるグラフィカルユーザインターフェイス、ウェブブラウザもしくはアプリを有するクライアントコンピュータといったフロントエンドコンポーネントを含む計算システムにおいて実現され得るか、または１つ以上のそのようなバックエンドコンポーネント、ミドルウェアコンポーネントもしくはフロントエンドコンポーネントの任意の組合せの計算システムにおいて実現され得る。システムのコンポーネントは、たとえば通信ネットワークといったデジタルデータ通信の任意の形態または媒体によって相互接続され得る。通信ネットワークの例は、ローカルエリアネットワーク（ＬＡＮ）およびワイドエリアネットワーク（ＷＡＮ）、たとえばインターネットを含む。

【００５９】

計算システムはクライアントおよびサーバを含むことができる。クライアントとサーバとは一般に互いから遠隔にあり、典型的には通信ネットワークを通じて対話する。クライアントとサーバとの関係は、それぞれのコンピュータ上で実行されるとともに互いに対してクライアント－サーバ関係を有するコンピュータプログラムによって発生する。いくつかの実施形態では、サーバは、例えば、クライアントとして振る舞うユーザデバイスと対

話するユーザにデータを表示し、およびそのユーザからユーザ入力を受信する目的で、データ、例えば、ＨＴＭＬページをユーザデバイスに送信する。ユーザデバイスで生成されたデータ、例えば、ユーザ対話の結果は、ユーザデバイスからサーバで受信することができる。

【００６０】

実施形態１は、ニューラルネットワークをトレーニングするための専用ハードウェアチップであって、専用ハードウェアチップの計算動作を制御するように構成されたスカラープロセッサと、ベクトル処理ユニットの２次元配列を有するように構成されたベクトルプロセッサとを備え、ベクトル処理ユニットは、すべて、同じ命令を単一命令複数データ方式で実行し、ベクトルプロセッサのロードおよびストア命令を通して互いに通信し、専用ハードウェアチップはさらに、ベクトルプロセッサに結合され、乗算結果を得るために、少なくとも１つの２次元行列を第２の１次元ベクトルまたは２次元行列と乗算するように構成された行列乗算ユニットを備える。

【００６１】

実施形態２は、ベクトルプロセッサに高速のプライベートメモリを提供するように構成されたベクトルメモリをさらに備える、実施形態１の専用ハードウェアチップである。

【００６２】

実施形態３は、スカラープロセッサに高速のプライベートメモリを提供するように構成されたスカラーメモリをさらに備える、実施形態１または２の専用ハードウェアチップである。

【００６３】

実施形態４は、行列の転置演算を実行するように構成された転置ユニットをさらに備える、実施形態１～３のいずれか１つの専用ハードウェアチップである。

【００６４】

実施形態５は、ベクトルアレイの異なるレーン間において、数値上で削減を実行し、数値を置換するように構成された、削減および置換ユニットをさらに備える、実施形態１～４のいずれか１つの専用ハードウェアチップである。

【００６５】

実施形態６は、専用ハードウェアチップのデータを記憶するように構成された高帯域幅メモリをさらに備える、実施形態１～５のいずれか１つの専用ハードウェアチップである。

【００６６】

実施形態７は、疎計算コアをさらに含む、実施形態１～６のいずれか１つの専用ハードウェアチップである。

【００６７】

実施形態８は、インターフェイスと、専用ハードウェアチップ上のインターフェイスまたはリソースを他の専用ハードウェアチップまたはリソースに接続するチップ間相互接続とをさらに備える、実施形態１～７のいずれか１つの専用ハードウェアチップである。

【００６８】

実施形態９は、複数の高帯域幅メモリをさらに備え、チップ間相互接続は、インターフェイスおよび高帯域幅メモリを他の専用ハードウェアチップに接続する、実施形態１～８のいずれか１つの専用ハードウェアチップである。

【００６９】

実施形態１０は、インターフェイスは、ホストコンピュータへのホストインターフェイスである、実施形態１～９のいずれか１つの専用ハードウェアチップである。

【００７０】

実施形態１１は、インターフェイスは、ホストコンピュータのネットワークへの標準ネットワークインターフェイスである、実施形態１～１０のいずれか１つの専用ハードウェアチップである。

【００７１】

本明細書は多くの特定の実現例の詳細を含んでいるが、これらは如何なる発明の範囲ま

たは請求され得るものの範囲に対する限定としても解釈されるべきではなく、特定の発明の特定の実施形態に特有の特徴であり得る記載として解釈されるべきである。別個の実施形態の文脈で本明細書において記載されるある特徴は、単一の実施形態において組合せでも実現され得る。反対に、単一の実施形態の文脈において記載されるさまざまな特徴は、複数の実施形態において別々に、または任意の好適な部分的組合わせでも実現され得る。さらに、特徴は、ある組合せにおいて作用すると上で記載され、最初はそのように請求されていさえする場合もあるが、請求される組合せからの１つ以上の特徴はいくつかの場合には当該組合せから削除され得、請求される組合せは、部分的組合わせまたは部分的組合わせの変形例に向けられ得る。

【００７２】

同様に、動作が図においては特定の順に示されているが、そのような動作は、望ましい結果を達成するために、示された当該特定の順もしくは連続した順で実行される必要があると理解されるべきではなく、または、すべての示された動作が実行される必要があると理解されるべきではない。ある状況においては、マルチタスキングおよび並列処理が有利であり得る。さらに、上述の実施形態におけるさまざまなシステムモジュールおよびコンポーネントの分離は、すべての実施形態においてそのような分離を必要とすると理解されるべきではなく、記載されるプログラムコンポーネントおよびシステムは一般に単一のソフトウェア製品に統合され得るかまたは複数のソフトウェア製品にパッケージ化され得ることが理解されるべきである。

【００７３】

主題の特定の実施形態が記載された。他の実施形態は以下の請求の範囲内にある。たとえば、請求項において記載されるアクションは、異なる順で実行され得、それでも望ましい結果を達成し得る。一例として、添付の図において示されるプロセスは、望ましい結果を達成するために、示された特定の順または連続する順であることを必ずしも必要としない。ある場合においては、マルチタスキングおよび並列処理が有利であり得る。

【図面】

【図１】

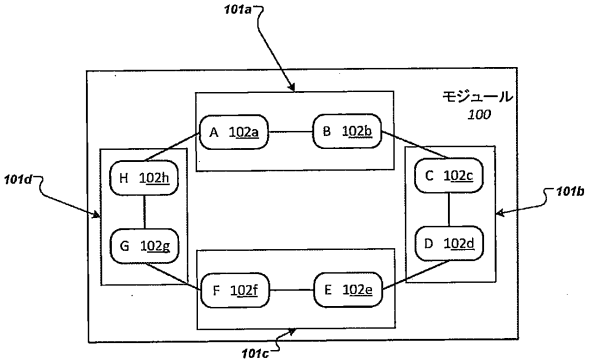

FIG. 1

【図２】

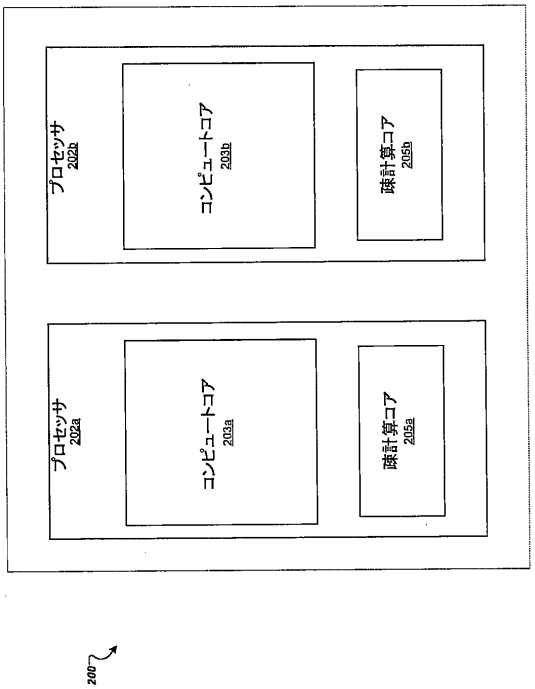

FIG. 2

【図３】

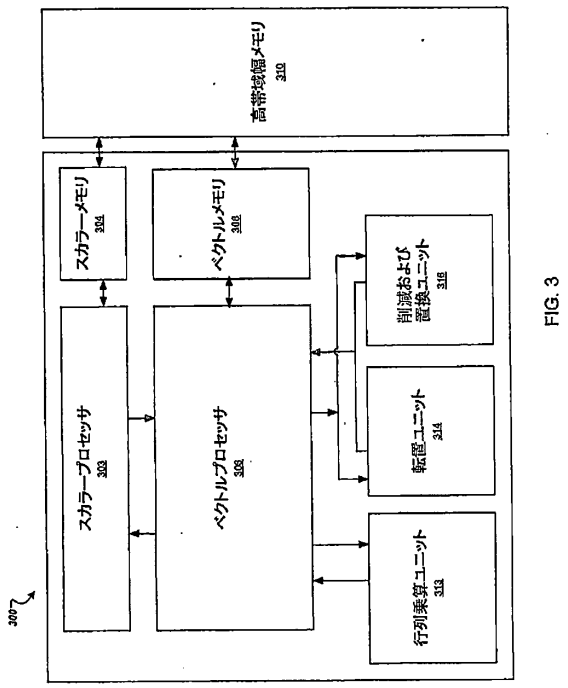

FIG. 3

【図４】

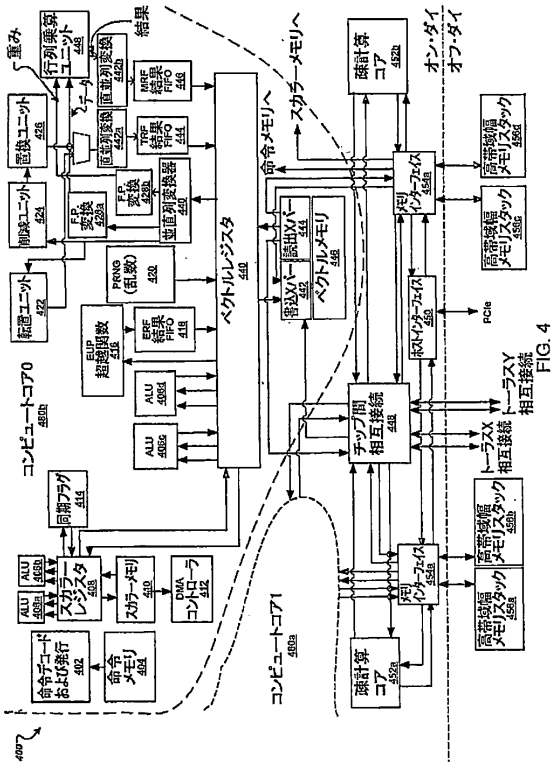

FIG. 4

フロントページの続き

(51)国際特許分類
G06F 17/10 (2006.01)
G06N 3/063(2023.01)

FI
G06F 17/16 P
G06F 17/10 S
G06N 3/063

(72)発明者 テマム，オリビエ
アメリカ合衆国、９４０４３ カリフォルニア州、マウンテン・ビュー、アンフィシアター・パークウェイ、１６００
(72)発明者 フェルプス，アンドリュー・エバレット
アメリカ合衆国、９４０４３ カリフォルニア州、マウンテン・ビュー、アンフィシアター・パークウェイ、１６００
(72)発明者 ジョピー，ノーマン・ポール
アメリカ合衆国、９４０４３ カリフォルニア州、マウンテン・ビュー、アンフィシアター・パークウェイ、１６００

審査官 坂東 博司

(56)参考文献
米国特許出願公開第２０１６／０３４２８９１（ＵＳ，Ａ１）
米国特許出願公開第２０１９／０３５４８６２（ＵＳ，Ａ１）
特開平０４－２９０１５５（ＪＰ，Ａ）
特開２０１７－１３８９６６（ＪＰ，Ａ）
下川 勝千，ディジタルニューロコンピュータＭＵＬＴＩＮＥＵＲＯ，東芝レビュー 第４６巻 第１２号，第46巻 第12号，日本，株式会社東芝，1991年12月，pp.931～934，【ISSN】0372-0462
斎藤 康毅，ゼロから作るＤｅｅｐ Ｌｅａｒｎｉｎｇ 初版，株式会社オライリー・ジャパン，第1版，日本，株式会社オライリー・ジャパン ，2016年09月，pp.123-165，第1版
ISBN: 978-4-87311-758-4

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ９／３８
Ｇ０６Ｆ ９／３０
Ｇ０６Ｆ １５／１７３
Ｇ０６Ｆ １５／８０
Ｇ０６Ｆ １７／１６
Ｇ０６Ｆ １７／１０
Ｇ０６Ｎ ３／０６３