【書類名】明細書

【発明の名称】ニューラルネットワークプロセッサ

【技術分野】

　本発明は、ニューラルネットワークプロセッサに関する。

【背景技術】

　ニューラルネットワークプロセッサは、ニューラルネットワークモデルの推論計算を実行するプロセッサである。ニューラルネットワークプロセッサは、例えばニューロンを模したコンピューティングチップであるニューロモーフィックプロセッサを含む。また、ニューロモーフィックプロセッサは、ニューロンを模した構成を有し、例えばスパイキングニューラルネットワークプロセッサとして提案されている。スパイキングニューラルネットワークプロセッサでは、ニューロン間を伝搬するスパイクが1ビットのデータに対応付けられる。

　本明細書で開示されるニューラルネットワークプロセッサは、ニューロモーフィックプロセッサや、スパイキングニューラルネットワークプロセッサに限定されない。但し、以下の実施の形態では、ニューラルネットワークプロセッサの一例であるニューロモーフィクプロセッサ（NMプロセッサと称する。）や、スパイキングニューラルネットワーク（以下SNNと称する。）プロセッサ等の実施の形態が開示される。

　ニューラルネットワークモデルは、入力データを受信する複数の第１のニューロン（ノード）と、複数の第１のニューロンとそれぞれのシナプス（エッジ）を介して接続された第２のニューロンとを有する。このような構成は、畳込演算層や全結合層と呼ばれる。そして、ニューラルネットワークモデルでは、複数の第１のニューロンの入力データとシナプスの重みとが積和演算され、第２のニューロンの活性化関数が積和演算結果に基づき算出する活性化出力が第２のニューロンの出力となる。シナプスの重みは、ニューラルネットワークの深層学習でファインチューニングされるパラメータの一つである。

　一方、SNNプロセッサでは、入力データや出力データとしてスパイクを模した１ビットのデータがニューロン間を伝搬する。そのため、SNNではニューロンによる演算が軽量化され、低消費電力化が可能であり、SNNは人工知能モデルのエッジコンピュータとして期待されている。

【先行技術文献】

　【特許文献】

　　　【特許文献１】米国特許公開公報2022/0129769号

　　　【特許文献２】米国特許第10,621,489号

　　　【特許文献３】米国特許第9,924,480号

　　　【特許文献４】米国特許第10,824,937号

　　　【特許文献５】特開平20-43541号公報

　　　【特許文献６】WO 2020/008642 A1

　【非特許文献】

　　　【非特許文献１】An In-Situ Dynamic Quantization With 3D Stacking Synaptic Memory for Power-Aware Neuromorphic Architecture, IEEE Access, Volume 11, 2023, August 3, 2023, NGO-DOANH NGUYEN, XUAN-TU TRAN, ABDERAZEK BEN ABDALLAH, and KHANH N. DANG

【発明の概要】

　　【発明が解決しようとする課題】

　ニューラルネットワークモデルやSNNモデルをプロセッサチップにマッピングすると、シナプスの重み等のパラメータを記憶するメモリがプロセッサチップ内の大きな面積を占有する。その結果、ニューロンにマッピングされた複数のコンピューティングコア間の距離が長くなるという課題がある。

　更に、メモリ内に記憶された重み等のパラメータにアクセスするときの電力消費が大きく、このパラメータのメモリの大きな消費電力は、エッジコンピュータを実現するために解決すべき課題である。そして、メモリのメモリセルに欠陥が発生した場合、特に重み等のパラメータの上位ビットを記憶するメモリセルに欠陥があると、ニューラルネットワークプロセッサによる演算精度の低下を招く可能性があることも課題である。

　そこで、本実施の形態の第1の側面の目的は、上記の課題を解決したニューラルネットワークプロセッサを提供することにある。

　　【課題を解決するための手段】

　本実施の形態の第１の側面は、複数のニューラルコンピューティングコアと、前記複数のニューラルコンピューティングコア間を接続するネットワークオンチップとを有する少なくとも１つのコンピューティングチップ層と、複数のパラメータをそれぞれ記憶する複数のメモリブロックをそれぞれする複数のメモリチップ層とを有し、前記複数のニューラルコンピューティングコアの各々は、前記ネットワークオンチップと接続されるネットワークインターフェースと、ニューロンモデルの演算を実行する複数の物理ニューロンを有するニューロンアレイを有し、前記パラメータは、第１のビット数を有する固定小数点フォーマットに量子化され、前記量子化されたパラメータは、少なくとも、上位ビットを有し前記第１のビット数より小さいビット数を有する第１の部分パラメータと、前記上位ビットより下位の下位ビットを有し前記第１のビット数より小さいビット数を有する第２の部分パラメータに分解され、前記積層された複数のメモリチップ層は、前記第１の部分パラメータを記憶する第１のメモリチップ層と、前記第２の部分パラメータを記憶する第２のメモリチップ層を有するニューラルネットワークプロセッサである。

　　【発明の効果】

　第１の側面によれば、上記の課題を解決することができる。

【図面の簡単な説明】

　【図１】ニューラルネットワークプロセッサ（以下NNプロセッサと称する。）の一例であるニューロモーフィックプロセッサ（NMプロセッサ）の構成例を示す図である。

　【図２】第1の実施の形態におけるNMプロセッサの側面図である。

　【図３】第1の実施の形態におけるNMプロセッサの各チップ層の構成例を示す分解斜視図である。

　【図４】第２の実施の形態におけるNMプロセッサの各チップ層の構成例を示す分解斜視図である。

　【図５】本実施の形態における重みの量子化プロセスと、量子化重みを部分重みに分割する分割プロセスを示す図である。

　【図６】動的量子化処理の様々なシナリオを示す図表である。

　【図７】本実施の形態における複数の部分重みを複数のメモリチップ層207\_1～207\_4に記憶した例を示す図である。

　【図８】電源制御方法の全体のフローチャートを示す図である。

　【図９】第1の電源制御方法における低電力モード605\_1でのメモリチップ層の電源制御のフローチャートを示す図である。

　【図１０】高性能モード604\_1でのメモリチップ層の電源制御のフローチャートを示す図である。

　【図１１】第２の電源制御方法における低電力モード605\_2でのメモリチップ層の電源制御のフローチャートを示す図である。

　【図１２】第２の電源制御方法における高性能モード604\_2でのメモリチップ層の電源制御のフローチャートを示す図である。

　【図１３】メモリチップ層の電源制御方法の供給電力と消費電力の変化を示す図である。

　【図１４】コンピューティングチップ層内の演算器の一例を示す図である。

　【図１５】図１４の演算器のマルチプレクサの動作例を示す図である。

　【図１６】欠陥メモリブロック内の部分重みの第１のリアロケート方法を示す図である。

　【図１７】欠陥メモリブロック内の部分重みの第２のリアロケート方法を示す図である。

　【図１８】第２のリアロケート方法の４つのパターンを示す図である。

　【図１９】第３の実施の形態におけるコンピューティングチップ層内の演算器の一例を示す図である。

　【図２０】図１９の加算器の入力段に設けられた回路の動作を説明する図である。

【発明を実施するための形態】

　ニューラルネットワークプロセッサは、入力データとニューラルネットワークの重み等のパラメータの畳み込み演算を実行する。ニューラルネットワークのパラメータには、重み以外にバイアスなど様々なデータが含まれる。本実施の形態では、パラメータの一つである重みを例にして、パラメータを記憶するメモリチップ層の構成とパラメータの処理について説明する。

　図１は、ニューラルネットワークプロセッサ（以下NNプロセッサと称する。）の一例であるニューロモーフィックプロセッサ（NMプロセッサ）の構成例を示す図である。図１に示したNMプロセッサ101は、半導体チップであり、マトリクス状に配置された複数のニューラルコンピューティングコア102と、コア間の通信を可能にする二次元メッシュ状のネットワークオンチップ103とを有する。NMプロセッサ101は、チップ間接続108を介してホストプロセッサチップ109に接続される。

　ニューラルコンピューティングコア102は、ネットワークインターフェース（NI）104と、重みメモリ（WM）105と、ニューロンアレイ（NA）106を有する。ネットワークインターフェース104は、ルータ107に接続され、ニューラルコンピューティングコア102とネットワークオンチップ103の間の通信のインターフェース処理を実行する。各コア102に設けられたルータ107は、ネットワークオンチップ103に設けられ、ネットワークオンチップ103を伝播する通信を宛先のコアの方向にルーティングする。更に、ルータ107は、コア宛の通信をそのコアの方向にルーティングする。

　前述のとおり、コアのネットワークインターフェース104が、ネットワークオンチップ103からルータ107によりルーティングされた入力を受信する。この入力は、例えば、ホストプロセッサ109が命令をデコードして発行した演算要求である。ネットワークオンチップ103は、受信した演算要求をデコードし、ニューロンアレイ106と重みメモリ105にデコードされた要求を送信する。それぞれの要求に基づいて、重みメモリ105は重みを読み出してニューロンアレイに送信し、ニューロンアレイ106は、要求に含まれる入力データと読み出された重みデータについて、要求に含まれた演算を実行する。演算結果は符号化されてルータ107に返信され、次の宛先のニューラルコンピューティングコア102に転送される。

　ニューロンアレイ106は、複数（Ｎ個）の物理ニューロン110を有し、識別番号SN1～SNNにインデックス付けされる。複数の物理ニューロン110は、ニューラルネットワーク内の複数のニューロンに対応付けられる。重みメモリ105は、複数のメモリブロック（WMB）111を有し、各メモリブロック111に記憶される重みデータは、ニューラルネットワーク内のニューロン間のエッジに対応付けられる。これらの対応付けはホストプロセッサ109により行われ、ホストプロセッサ109は、この対応付けに基づいて演算要求を発行し、演算要求の宛先のニューラルコンピューティングコア102に送信する。

各物理ニューロン110は、演算器を有し、要求された演算を実行する。物理ニューロン110は、ネットワークインターフェース104から入力データ等を含む要求を受信し、重みメモリ105から演算要求に基づいて読み出された重みデータを受信する。そして物理ニューロン110は、入力データと重みデータを使用して要求された演算を実行する。

　要求された演算が畳み込み演算の場合、演算を実行する物理ニューロン110は、入力ニューロンから送信された入力データと、入力ニューロンのエッジに対応する重みデータを受信する。そして、演算を実行する物理ニューロン110は、入力データと重みデータの積和演算を実行し、ニューロンモデルの活性化関数に積和演算結果を入力して実行し、その演算結果を出力する。物理ニューロンの演算は、例えばNMプロセッサ内のクロックサイクルに同期して実行され、演算結果は次の演算実行対象の物理ニューロンに送信される。

即ち、Ｍ個の入力ニューロンから送信された入力データをinput1～inputM、各エッジの重みをweight1～weightM、活性化関数をfactivationとすると、演算を実行する物理ニューロンの出力outputは次のとおりである。  
　output = factivation (Σ (i=1-M) inputi \* weighti)  
ここで、Ｍは入力データの数、重みの数である。活性化関数はReLU関数などである。

　［第1の実施の形態］  
　図２は、第1の実施の形態におけるNMプロセッサの側面図である。図３は、NMプロセッサの各チップ層の構成例を示す分解斜視図である。図２、図３に示すとおり、NMプロセッサ213は、複数のチップ層201、207で構成される。複数のチップ層は、複数のニューラルコンピューティングコア205とネットワークオンチップ204を有するコンピューティングチップ層201を有する。更に、複数のチップ層は、コンピューティングチップ層201上に積み重ねられた複数のメモリチップ層207を有する。

　複数のチップ層201、207は、パッケージ基板215上に実装される。また、複数のチップ層201、207は、層間ビア209を介して通信可能にされる。複数のチップ層201、207内に設けられた層間通信部208が、層間ビア209を介して複数チップ間の通信を行う。複数のメモリチップ層207は、例えばSRAMやReRAM等の高速メモリを有する。

　コンピューティングチップ層201内のネットワークオンチップ204は、チップ間接続214を介して、ホストプロセッサ211に接続される。また、ネットワークオンチップ204とルータ202は、複数のニューラルコンピューティングコア205間の通信を可能にする。ルータ202は、各コア205に対応して設けられる。

　コンピューティングチップ層201内のニューラルコンピューティングコア205は、ネットワークインターフェース（NI）203と、ニューロンアレイ（NA）212と、層間通信部208を有する。しかし、コア205は重みメモリを有していない。

　一方、複数のメモリチップ層207はそれぞれ、重みやその他のパラメータを記憶する重みメモリ206を有する。そのため、コンピューティングチップ層201内のニューラルコンピューティングコア205は、重みメモリを有していない。したがって、コアの面積が図１の例よりも小さくなる。これにより、ネットワークオンチップ204上に配置される複数のルータ202間の距離が短くなる。このことは、コア205間の通信時間が短くなることを意味する。また、より多くのコア205をコンピューティングチップ層201に設けることができる。

　本実施の形態ではNMプロセッサは、複数の重みやその他のパラメータをメモリチップ層207内の重みメモリ206に記憶する。また、浮動小数点の重みが第1のビット数の固定小数点フォーマットで量子化され、量子化重みが重みメモリ206に記憶される。これに伴い、物理ニューロン内の演算器は、固定小数点演算器であり、固定小数点の重みを入力して 演算する。

　複数のメモリチップ層207\_1～207\_Nは、量子化された重みをＮ個に分割して生成された部分重みを、それぞれ記憶する。即ち、第1のビット数の量子化重みが、第1のビット数より少ない第２のビット数の第１乃至第Ｎの部分重みに分解される。図２の例ではＮ＝４である。そして、最上位ビットMSBを有する第１の部分重みは、第１のメモリチップ層207\_1内の重みメモリに記憶され、最下位ビットLSBを有する第Ｎの部分重みは、第Ｎのメモリチップ層207\_N内の重みメモリに記憶される。第１から第Ｎの部分重みは、順に部分重みのビットの桁が下がるように分割され、第１から第Ｎのメモリチップ層にそれぞれ記憶される。分割された部分重みは、必ずしも同じビット数（第２のビット数）である必要はなく、異なるビット数であってもよい。

　上記のように複数の部分重みを第１から第Ｎのメモリチップ層に別々に記憶すると、第１のメモリチップ層にはMSBを含む部分重みが、第２乃至第Ｎ－１のメモリチップ層には桁が順々に下がる部分重みが、そして、第Ｎのメモリチップ層にはLSBを含む部分重みが記憶される。４層のメモリチップ層の具体的が後で図５乃至７に示される。

　第１の実施の形態によれば、第1に、Ｎ個のメモリチップ層207に、上位ビットの部分重みから下位ビットの部分重みを別々に記憶するので、下位ビットの部分重みを記憶するメモリチップ層の電源をオフにして省電力化できる。その場合、電源オフのメモリチップ層が記憶する下位ビットの部分重みは使用できなくなるが、下位ビットであるのでNMプロセッサの精度低下を抑制できる。

　第２に、欠陥が発生したメモリブロック内の上位ビットの部分重みを、下位ビットの部分重みを記憶する未欠陥のメモリブロックに上書きして、重みメモリのフォールトトレランスを向上させることができる。この場合下位ビットの重みが使用できなくなるが、下位ビットであるのでNMプロセッサの精度低下を抑制できる。

　［第２の実施の形態］

　図４は、第２の実施の形態におけるNMプロセッサの各チップ層の構成例を示す分解斜視図である。図４のNMプロセッサ213は、図２、図３に示したNMプロセッサと同様の構成である。即ち、NMプロセッサ213は、コンピューティングチップ層201と、第１乃至第Ｎのメモリチップ層207\_1～207\_Nとを積層する。コンピューティングチップ層201は、複数のニューラルコンピューティングコア205と、その間の通信に使用されるネットワークオンチップ204を有する。そして、Ｎ層のメモリチップ層207\_1～207\_N内に、重み及びその他のパラメータが記憶される。

　図３と異なり、図４では、複数のメモリチップ層207\_1～207\_Nに、電源レール1505と、電源制御回路1502と、電圧レギュレータ1503が設けられる。また、コンピューティングチップ層201に、電源レール1505と、電源入出力ポート1500が設けられる。そして、これらチップ層207、201の電源レール1505が電源ビア1501を介して接続される。電源入出力ポート1500が外部の電源に接続され、電力がNMプロセッサ213に供給される。外部電源から供給される電力は、電源レール、電源ビア等を経由して各チップ層201、207の電子回路に供給される。図４内の電源レール、電源ビア、電源制御回路、電圧レギュレータの構成が、図３とは異なる。

　ホストプロセッサ211からの電源制御信号（図示せず）に基づいて、各メモリチップ層207\_1～207\_N内の電源制御回路1502が、メモリチップ層内の電源のオン、オフ制御を行う。電源制御回路1502が電源オン制御すると、その電源が電圧レギュレータ1503に供給され、電圧レギュレータ1503は、ホストプロセッサ211からの電圧制御信号に基づいて、電源レール1505の電源電圧を上昇または下降制御する。電圧レギュレータにより電圧制御された電源電圧が、電源レール1505を介して、メモリチップ層内の各回路モジュールに供給される。

　図５は、本実施の形態における重みの量子化プロセスと、量子化重みを部分重みに分割する分割プロセスを示す図である。図５の上部には、ニューラルネットワークモデルの構成例が示される。この例では、ニューラルネットワークモデル301は、入力側のニューロンモデル302\_INと、出力側のニューロンモデル302\_OUTと、両ニューロンモデル間のエッジの重みモデル303とを有する。重みモデル303は、通常、浮動小数点フォーマットの重みWik、Wjkである。重みのサフィックスi, jは、入力側ニューロンモデルiと出力側ニューロンモデルｊである。

　量子化プロセス304により、浮動小数点の重み303は、固定数ビット、例えば８ビットの固定小数点の重みWik [0:7]、Wjk [0:7]に量子化される。量子化重み305は、８ビットの固定小数点フォーマットで量子化さる。尚、[0:7]の０側がMSB、７側がLSBのビットに対応する。

　分解プロセス306により、８ビットの量子化重みWik [0:7]、Wjk [0:7]が、例えば４つの部分重み307に分解される。例えば、部分重みは２ビットである。以下具体的を示す。  
第1の部分重み307\_1は、Wik [0:1]、Wjk [0:1]で、MSBと２桁目を含む２ビットの部分重みである。  
第２の部分重み307\_2は、Wik [2:3]、Wjk [2:3]で、３、４桁目を含む２ビットの部分重みである。  
第３の部分重み307\_3は、Wik [4:5]、Wjk [4:7]で、５，６桁目を含む２ビットの部分重みである。  
第4の部分重み307\_4は、Wik [6:7]、Wjk [6:7]で、７桁目とLSBを含む２ビットの部分重みである。

　図６は、動的量子化処理の様々なシナリオを示す図表である。図中、列401は浮動小数点の重みモデル303の例を、列402は重みモデル303の２進数を、列403は８ビットの量子化重み305を、列404は量子化重み305の１０進数を、列405は２ビットずつに分解された部分重み307の例をそれぞれ示す。

　更に、列406は最下位の部分重みのビットが反転した場合の部分重み307を示す。反転ビットには下線が付されている。列407は反転したビットの値を１０進数で示す。これによれば、最下位の部分重みのビットが反転しても、重みの値の変化は限定的であることが分かる。

　図７は、本実施の形態における複数の部分重みを複数のメモリチップ層207\_1～207\_4に記憶した例を示す図である。図中、コンピューティングチップ層201と、４つのメモリチップ層207\_1～207\_4とが積層されている。但し、上層の２つのメモリチップ層207\_3、207\_4は、右側に並べて示される。各メモリチップ層207\_1～207\_4それぞれの重みメモリ206\_1～206\_4のメモリブロックMB\_1～MB\_4には、４つの部分重みWik [0:1]～Wik [6:7]が別々に記憶される。

　好ましくは、ニューロンi,k間の部分重みWik [0:1]、Wik [2:3]、Wik [4:5]、Wik [6:7]はそれぞれ、重みメモリ206\_1、206\_2、206\_3、206\_4内の同じチップ内アドレスのメモリブロックMB\_1～MB\_4に記憶される。ニューロンj,k間の部分重みWjk [0:1]、Wjk [2:3]、Wjk [4:5]、Wjk [6:7]も同様に同じチップ内アドレスのメモリブロックに記憶される。

このように４つの部分重みを記憶するメモリブロックをメモリチップ層内の同じアドレスにマッピングすることで、４つの部分重みの読み出し処理が簡素化される。つまり、ホストプロセッサ211は、重みWik［0：7]のチップ内アドレスを指定することで、４つの部分重みWik [0:1]、Wik [2:3]、Wik [4:5]、Wik [6:7]を、４層のメモリチップ層207からそれぞれ読み出すことができる。そして、物理ニューロンの演算器は、読み出した４つの部分重みのビットをマージすることで、８ビットの重みWik［0：7]を取得することができる。そして、マージした８ビットの重みが固定小数点演算器に入力される。但し、ある重みの複数の部分重みを同じチップ内アドレスのメモリブロックに記憶しなくてもよい。その場合、演算命令は４つの部分重みそれぞれのチップ内アドレスを含み、コア内のネットワークインターフェースは、それぞれのチップ内アドレスで部分重みを記憶するメモリブロックにアクセスする。

　更に、好ましくは、第１の部分重みW［0:1］は、コンピューティングチップ層201に最も近い第1のメモリチップ層207\_1内の重みメモリ206\_1内に格納される。この理由は、量子化重みW［0：7］のうち最上位ビットMSBを含む第1の部分重みW［0:1］が、ニューラルネットワークプロセッサの推論の精度に最も大きな影響を与えるからである。コンピューティングチップ層201は、MSBの第1の部分重みを、最も近い第1のメモリチップ層207\_1から高速に読み出すことができる。

　［電源制御方法］

　第２の実施の形態では、NMプロセッサの供給電力より消費電力が高くなると、複数のメモリチップ層の消費電力を低下させる電源制御を行う。以下、２つの電源制御方法について説明する。

電源制御方法では、ホストプロセッサ211がメモリチップ層内の電源制御回路1502と、電圧レギュレータ1503に電源制御信号を送信して、メモリチップ層の電源制御を実行する。図４のNMプロセッサにおいて、ホストプロセッサ211が発行した電源制御信号が、チップ間接続214と、コンピューティングチップ層201と、層間ビア209を経由して、制御先のメモリチップ層の電源制御回路1502に送信される。

　図８は、電源制御方法の全体のフローチャートを示す図である。電源制御600を行うホストプロセッサ211は、NMプロセッサの供給電力PsupplyがNMプロセッサの消費電力Pconsumedより大きい場合（601のYES）、メモリチップ層を高性能モードで制御する（604）。一方、ホストプロセッサ211は、NMプロセッサの供給電力PsupplyがNMプロセッサの消費電力Pconsumedより小さい場合（602のYES）、メモリチップ層を低電力モードで制御する（605）。そして、ホストプロセッサ211は、NMプロセッサの供給電力PsupplyがNMプロセッサの消費電力Pconsumedと等しい場合（601、602のNO）、現在の制御モードを維持する（603）。上記の電源制御が所定時間T毎に実行される。

　［第1の電源制御方法（動的量子化による電源制御）］

　図９は、第1の電源制御方法における低電力モード605\_1でのメモリチップ層の電源制御のフローチャートを示す図である。フローチャートの右下に、４つのメモリチップ層の電源オン、オフの制御が時間軸に対応して示される。

　ホストプロセッサ211は、電源がアクティブ状態（オン状態）の最上層のメモリチップ層のインデックスkから電源制御を開始する（701）。全てのメモリチップ層が電源オン状態であれば、ホストプロセッサは、最上層の第４のメモリチップ層から電源制御を行う。NMプロセッサへの供給電力PsupplyがNMプロセッサの消費電力Pconsumedより小さいまたは以下の場合（702のYES）、低電力モード605に入る。低電力モードでは、インデックスkがメモリチップ層の最下層のインデックス“１”に達していない場合（703のNO）、そのメモリチップ層kの電源がオフに制御される（704）。そして、インデックスkが１つ減らされ（705）、前述と同じ低電力モード処理（702、703、704）が行われる。インデックスkが最下層のインデックス“１”に達すると（703のYES）、部分重みのMSBを格納する最下層のメモリチップ層の電源をオフにできないので（706）、低電力モードが終了する。

　フローチャートの右下に示したとおり、全てのメモリチップ層の電源がオンの状態（k＝４）で低電力モードに入ると、NMプロセッサの供給電力PsupplyがNMプロセッサの消費電力Pconsumedより小さいまたは以下の間は、最上層のメモリチップ層（k＝４）から最下層より一つ上の層（k＝２）のメモリチップ層に向かって順に電源オフに制御される。NMプロセッサは、第1のメモリチップ層より一つ上の第２のメモリチップ層（k＝２）まで電源がオフに制御されると、電源オンの第1のメモリチップ層に記憶されるMSBの部分重みを使用して推論を継続する。この場合、推論精度は低下するが、MSBを含む第１の部分重みで演算されるため、推論精度の低下は限定的である。

　図１０は、高性能モード604\_1でのメモリチップ層の電源制御のフローチャートを示す図である。ホストプロセッサ211は、電源がアクティブ状態（オン状態）の最上層のメモリチップ層のインデックスkから開始し（801）、NMプロセッサの供給電力PsupplyがNMプロセッサの消費電力Pconsumedより大きいまたは以上の場合（802のYES）、高性能モード604\_1に入る。

　高性能モードでは、インデックスkがメモリチップ層の最上層のＮに達していない間（803のNO）、インデックスkが１つ増やされ（804）、メモリチップ層kが電源オンに制御される（805）。電源オンのメモリチップ層が増えたため、NMプロセッサの供給電力PsupplyがNMプロセッサの消費電力Pconsumedより小さくなったかがチェックされる（806）。この比較806がNOの場合、前述と同じ高性能モード処理（802～805）が行われる。この比較806がYESの場合、直前に電源オンに制御されたインデックスｋのメモリチップ層の電源はオフに戻されて（807）、高性能モードが終了する。また、インデックスkが最上層Ｎに達した場合（803のYES）も、高性能モードが終了する。

　フローチャートの右下に示したとおり、最上層から第２層のメモリチップ層が電源オフの状態で高性能モードに入ると、NMプロセッサの供給電力PsupplyがNMプロセッサの消費電力Pconsumedより大きい間は、第２層のメモリチップ層から最上層のメモリチップ層に向かって順に電源オンに制御される。

　上記の図９の低電力モード605\_1に入ると、下位ビットの部分重みが順に利用できなくなり、利用できる量子化重みのビット数が低下する。一方、上記の図１０の高性能モード604\_1に入ると、下位ビットの部分重みが順に利用できるようになり、利用できる量子化重みのビット数が増加する。したがって、量子化重みのビット数が動的に変更制御される。

　図９において、制御開始時のメモリチップ層のインデックスkは、必ずしも最上層の第４のメモリチップ層（k＝４）になるとは限らない。電源オン状態の最上層が第３のメモリチップ層の場合、開始時のインデックスkは３となる。同様にk＝２から開始する場合もある。逆に、図１０において、開始時のメモリチップ層のインデックスkは、必ずしも最下層の第１のメモリチップ層（k＝１）になるとは限らない、電源オン状態の最上層が第２のメモリチップ層の場合、開始時のインデックスkは２となる。同様にk＝３から開始する場合もある。

　［第２の電源制御方法（動的量子化と電圧スケーリングによる電源制御）］

　第２の電源制御方法では、低電力モードにおいて、ホストプロセッサ211は、メモリチップ層のアクティブな電源をオフにする前に、そのメモリチップ層の電源電圧を低下する。そして、ホストプロセッサは、最小電圧Vminに達しても供給電力が消費電力より小さいまたは以下なら、そのメモリチップ層の電源をオフにする。

逆に、高性能モードにおいて、ホストプロセッサは、メモリチップ層の非アクティブな電源をオンにする前に、一つ下層のメモリチップ層の電源電圧を上昇する。そして、ホストプロセッサは、最大電圧Vmaxに達しても供給電力が消費電力より大きいまたは以上なら、一つ上層のメモリチップ層の電源をオンにする。第２の電源制御方法の電源電圧の上昇または低下（スケーリング）以外の動作は、第1の電源制御方法と同じである。

　図１１は、第２の電源制御方法における低電力モード605\_2でのメモリチップ層の電源制御のフローチャートを示す図である。このフローチャートでは、ホストプロセッサ211は、制御対象のインデックスkの電源電圧Vkが、動作可能な最小電圧Vminに達していなければ（708のNO）、その電源電圧Vkを一定量降下させる（709）。ホストプロセッサは、電源電圧Vkが動作可能な最小電圧Vminに達していれば（708のNO）、kが１ではない場合（703のNO）、インデックスkのメモリチップ層を電源オフに制御する（704、705）。そして、ホストプロセッサ211は、この処理708と709以外は、図９の第1の電源制御方法における低電力モード605\_1と同じ制御を実行する。

　図１２は、第２の電源制御方法における高性能モード604\_2でのメモリチップ層の電源制御のフローチャートを示す図である。このフローチャートでは、ホストプロセッサ211は、制御対象のインデックスkの電源電圧Vkが最大電圧Vmaxに達していなければ（808のNO）、その電源電圧Vkを一定量上昇させる（809）。ホストプロセッサは、電源電圧Vkが最大電圧Vmaxに達していれば（808のNO）、kがNではない場合（803のNO）、一つ上層のインデックスk+1のメモリチップ層を電源オンに制御する（804～807）。そして、ホストプロセッサ211は、この処理808と809以外は、図１０の第1の電源制御方法における高性能モード604\_1と同じ制御を実行する。

　図１３は、メモリチップ層の電源制御方法の供給電力と消費電力の変化を示す図である。図１３の上段の605\_1、604\_1は、第１の電源制御方法の動的量子化のみの電力変化例である。図１３の下段の605\_2、604\_2は、第２の電源制御方法の動的量子化及び電源電圧スケーリングの電力変化例である。

図１３の上段の605\_1、604\_1では、メモリチップ層の電源オン・オフ制御が行われ、消費電力の変化が大きな階段状となる。図１３の下段の605\_2、604\_2では、メモリチップ層の電源オン・オフに加えてメモリチップ層内の電源電圧の上昇・下降制御が行われ、消費電力の変化が小さな階段状となる。また、高性能モード604\_1、604\_2では、フローチャートのステップ704、805により消費電力が一旦上昇した後に低下する制御が行われる。

　図１４は、コンピューティングチップ層内の演算器の一例を示す図である。演算器である加算器ADDERは、重み入力in１と、入力ノードの入力in２とを入力し、加算演算を実行し、演算結果outを出力する。加算器の重み入力in1は８ビットの固定小数点数である。本実施の形態では、メモリチップ層の電源制御が行われ、NMプロセッサが高性能モードまたは省電力モードの動作に制御される。高性能電力モードでは、例えば全ての部分重みが重みとして利用され、一方、省電力モードでは例えばMSBの部分重みのみが重みとして利用される。

　電源制御により部分重みメモリから読み出される部分重みの数が変動した場合、加算器は、マルチプレクサMUX11～MUX14を介して、メモリから読み出された有効な部分重みと、無効な部分重みの代わりの２ビット信号00とを入力する。図中、pw1～pw4はメモリチップ層から読み出される部分重みである。on1～on4は、メモリチップ層の電源オン・オフに対応した制御信号である。

　メモリチップ層から読み出される部分重みpw1～pw4は、電源オンのメモリチップ層から読み出された有効なビットと、電源オフのメモリチップ層から読み出された無効なビットとを含む。マルチプレクサは、電源オン・オフに対応した制御信号on1～on4に基づいて、無効な部分重みビットpw1～pw4を２ビット信号00に置き換えて出力し（制御信号on1～on4が0の場合）、または有効な部分重みビットpw1～pw4をそのまま出力する（制御信号on1～on4が1の場合）。

　図１５は、図１４の演算器のマルチプレクサの動作例を示す図である。前提として、４つの部分重みpw1～pw4はいずれも２ビット０１とする。読み出された４つの部分重みpw1～pw4は、無効ビットの場合ハイインピーダンスzとなる。  
　図１５の例では、時間timeの経過に伴って、メモリチップ層の電源が下位ビット側の層から順にオフになる。左端のクロックタイミングCK1では、全制御信号on1～on4が“１”であるため、全マルチプレクサはメモリチップ層から読み出した有効な部分重みpw1～pw4をそのまま出力する。そして、クロックタイミングがCK2～CK4に移動すると、下位ビットの部分重みから順に無効ビットzとなり、２ビット信号００に置き換えられる。演算器に入力される重み入力in1の８ビットのうち、下線部のビット００が置き換えられたことを示す。

　以上の通り、第２の実施の形態では、複数の部分重みが複数のメモリチップ層にそれぞれ記憶されるので、省電力モードでは、ホストプロセッサが、最下位ビットの部分重みのメモリチップ層から順に電源オフに制御する。また、高性能電力モードでは、ホストプロセッサが、２番目に高いビットの部分重みのメモリチップ層から順に電源オンに制御する。これにより、ホストプロセッサは、NMプロセッサの推論精度の低下を抑制しつつ、省電力モードと高性能電力モードの電源制御を行うことができる。

［第３の実施の形態］

　第３の実施の形態では、メモリチップ層内のあるメモリブロックが欠陥になり、そこに記憶された部分重みpwが欠陥になった場合、欠陥メモリブロック内の部分重みpwを、下位ビットの部分重みを記憶する未欠陥メモリブロックにリアロケート（再割り当て）する。これにより、NMプロセッサの演算精度への影響が大きい上位ビットの部分重みの利用を継続することができる。欠陥メモリブロック内の部分重みのリアロケート方法について以下説明する。

　［第1のリアロケート方法］

　図１６は、欠陥メモリブロック内の部分重みの第１のリアロケート方法を示す図である。図１６には、コンピューティングチップ層201と、４つのメモリチップ層207\_1～207\_4とが示される。仮に、最上位ビットを含む部分重みWik[0:1]を記憶するメモリブロックMB\_1に欠陥があり、そこから読み出された部分重みWik[0:1]にエラービットFaulty pwが検出されたとする。

ホストプロセッサ211は、エラービットを有する部分重みをリストアップする。ホストプロセッサは、エラービットを有する部分重みがLSBの部分重みを記憶するメモリチップ層のメモリブロック内か否かをチェックする。チェック結果が偽であれば、ホストプロセッサ211は、欠陥メモリブロックが記憶する部分重みWik[0:1]を、その最下位ビットを含む部分重みWik[6:7]を記憶するメモリブロックMB\_4に上書きして、部分重みWik[0:1]をリアロケートReAする。これにより、最下位ビットを含む部分重みWik[6:7]の利用は停止するが、最上位ビットを含む部分重みWik[0:1]の利用は継続できる。

　チェック結果が真であれば、ホストプロセッサ211は、エラービットの部分重みをリアロケートせずに、その最下位ビットを含む部分重みの利用を停止する。

　MSBを含む部分重みWik[0:1]が、メモリチップ層207\_1からメモリチップ層207\_4にリアロケートされた結果、ホストプロセッサ211は、NMプロセッサに発行する演算要求内の部分重みWik[0:1]のアドレスをリアロケート先のアドレスに変換する。このアドレス変換に基づき、NMプロセッサのネットワークインターフェースNIは、演算要求内のリアロケート先のアドレスに基づいて部分重みWik[0:1]をメモリチップ層207\_4から読み出す。

　［第２のリアロケート方法］

　図１７は、欠陥メモリブロック内の部分重みの第２のリアロケート方法を示す図である。図１７は、図１６と同様に、１つのコンピューティングチップ層201と、４つのメモリチップ層207\_1～207\_4を有し、部分重みWik[0:1]～Wik[6:7]、Wjk[0:1]～Wjk[6:7]も同様に記憶される。そして、最上位ビットを含む部分重みWik[0:1]を記憶するメモリブロックMB\_1に欠陥があったことも同様である。

前提として、各４つの部分重みWik[0:1]～Wik[6:7]、Wjk[0:1]～Wjk[6:7]は、４つのメモリチップ層内のそれぞれの同じアドレスのメモリブロックに記憶される。

　第２のリアロケート方法では、ホストプロセッサ211は、エラービットを有する部分重みをリストアップする。ホストプロセッサは、エラービットを有する部分重みがLSBの部分重みを記憶するメモリチップ層のメモリブロック内か否かをチェックする。

チェック結果が偽であれば、ホストプロセッサ211は、メモリチップ層207\_1、207\_2、207\_3内の部分重みWik[0:1]、Wik[2:3] 、Wik[4:5]を、１つ上層のメモリチップ層207\_2、207\_3、207\_4内のメモリブロックにリアロケートする。図中リアロケートReAに示すとおりである。つまり、欠陥が発生した部分重みWik[0:1]からLSBの部分重みWik[6:7]より１つ上位ビットの部分重みWik[4:5]までの部分重みの束を、それぞれ１つ上層のメモリチップ層内のメモリブロックにリアロケートする。このようにリアロケートした場合、ホストプロセッサ211は、演算要求内の部分重みのアドレスを変更せず、演算器の入力段のマルチプレクサへの制御信号を変更する。後述するとおり、制御信号を変更することで、リアロケートされた部分重みの束は、元のメモリチップ層の出力として演算器に入力される。よって、ホストプロセッサ211の負荷を軽減できる。

　チェック結果が真であれば、ホストプロセッサ211は、エラービットの部分重みをリアロケートせずに、その最下位ビットを含む部分重みの利用を停止する。

　図１８は、第２のリアロケート方法の４つのパターンを示す図である。各表の５つの列は、メモリチップ層Layer、未欠陥か欠陥T/F、リアロケート前の各メモリチップ層に記憶された部分重みoriginal、リアロケート後のメモリチップ層に記憶された部分重みrealct、そして、演算器の重み入力in1の部分重みW[0:1]～W[6:7]を示すin1である。図１９で示す演算器の入力段のマルチプレクサにより、メモリチップ層から読み出されたリアロケート後の部分重みが、演算器の入力in1のビットに対応付けられる。

　第１のリアロケートパターン1611は、図１７に示した例である。つまり、最上位ビットの部分重みpw1を記憶するメモリチップ層のメモリブロックが欠陥Fになり、欠陥メモリブロックの部分重みw[0:1]から部分重みw[2:3]、w[4:5]までの束を、それぞれ１つ上層のメモリチップ層pw2～pw4にリアロケートした例である。ホストプロセッサ211は、演算要求内の重みのアドレスを変更せず、リアロケートされた部分重みw[0:1]、w[2:3]、w[4:5]を、それぞれ１つ上のメモリチップ層から読み出す。そして、読み出された部分重みは、後述する演算器の入力段のマルチプレクサによりリアロケート前の部分重みの元のビット位置にシフトされ、演算器に入力される。

　第２のリアロケートパターン1612は、最上位ビットから２番目のメモリチップ層の部分重みpw2を記憶するメモリブロックが欠陥Fになり、欠陥メモリブロックの部分重みw[2:3]から部分重みw[4:5]までの束を、それぞれ１つ上層のメモリチップ層pw3～pw4にリアロケートした例である。この場合も、ホストプロセッサ211は、演算要求内の重みのアドレスを変更せず、リアロケートされた部分重みw[2:3]、w[4:5]を、それぞれ１つ上のメモリチップ層から読み出す。そして、読み出された部分重みは、後述する演算器の入力段のマルチプレクサによりリアロケート前の部分重みのビット位置にシフトされ、演算器に入力される。

　第３のリアロケートパターン1613は、最上位ビットから３番目のメモリチップ層の部分重みpw3を記憶するメモリブロックが欠陥Fになり、欠陥メモリブロックの部分重みw[4:5]を、１つ上層のメモリチップ層pw4にリアロケートした例である。この場合も、ホストプロセッサ211は、演算要求内の重みのアドレスを変更せず、リアロケートされた部分重みw[4:5]を、それぞれ１つ上のメモリチップ層から読み出す。そして、読み出された部分重みw[4:5]は、後述する演算器の入力段のマルチプレクサによりリアロケート前の部分重みの元のビット位置にシフトされ、演算器に入力される。

　最後に、第４のリアロケートパターン1614は、最下位のメモリチップ層の部分重みpw4を記憶するメモリブロックが欠陥Fになり、リアロケートされない例である。欠陥メモリブロックの部分重みw[6:7]は、入力段のマルチプレクサにより固定の２ビット00に変換され、演算器に入力される。

　図１９は、第３の実施の形態におけるコンピューティングチップ層内の演算器の一例を示す図である。図１４と同様に、演算器である加算器ADDERは、重み入力in１と、入力ノードの入力in２とを入力し、加算演算を実行し、演算結果outを出力する。重み入力in1は８ビットの固定小数点フォーマットのデータである。

　本実施の形態では、メモリチップ層内の欠陥メモリブロック内の部分重みを、他のメモリチップ内の未欠陥メモリブロックにリアロケートする。そのため、４つのメモリチップ層から読み出される部分重みは、リアロケート先のメモリチップ層から読み出される。

　図１９中、pw1～pw4はメモリチップ層207\_1～207\_4から読み出される部分重みである。制御信号on1～on4は、メモリチップ層の部分重みの無効・有効に対応する。制御信号s1～S3は、メモリチップ層から読み出された部分重みpw2～pw4のシフトの有無に対応する。

図１９の加算器では、リアロケート先のメモリチップ層から読み出された部分重みpw１～pw4を、マルチプレクサMUX1～MUX3により、重みの元のビット位置にシフトダウンする。これにより、リアロケートされた部分重みビットは、８ビットの重みの元のビットとして加算器に入力される。また、別のマルチプレクサMUX11～MUX14により、無効の部分重みを固定の２ビット00に変換する。これにより、無効の部分重みビットは00に変換され、加算器に入力される。

　図２０は、図１９の加算器の入力段に設けられた回路の動作を説明する図である。クロックClockに示したクロックタイミングCK1～CK4での、４つのメモリチップ層から読み出した部分重みpw1～pw4と、制御信号on1～on4と、制御信号s1～S3が示される。

　クロックCK1に、全てのメモリチップ層に欠陥がない場合の信号が示される。メモリチップ層から読み出される部分重みpw1～pw4は一例として「01」「10」「01」「10」である。欠陥がないので、ホストプロセッサ211から送信された制御信号on1～on4、s1～S3は全て0である。

　クロックCK2に、最下層のメモリチップ層内の最上位ビットの部分重みのメモリブロックに欠陥が発生した場合の信号が示される。この場合、ホストプロセッサ211が、部分重みw[0:1]、w[2:3]、w[4:5]の束を、それぞれ１つ上層のメモリチップ層pw2～pw4にリアロケートする。つまり、前述の第１のリアロケートパターン1611に対応する。

　この場合、ホストプロセッサ211は、制御信号s1～S3を「111」とする。この制御信号に基づき、３つのマルチプレクサMUX1～MUX3は、リアロケート先のメモリチップ層から読み出した部分重みpw2～pw4を重みの元のビット位置にシフトダウンする。更に、ホストプロセッサは、制御信号on1～on4を「0001」とし、マルチプレクサMUX14に固定ビット00を出力させる。残りのマルチプレクサMUX11～MUX13は、シフトダウンした部分重みを出力する。この結果、加算器の入力in1は、「01100100」となる。

　次に、クロックCK3に全てのメモリチップ層に欠陥がない場合の信号が示される。クロックCK1と同じである。そして、クロックCK4に、第３層のメモリチップ層のメモリブロックに欠陥が発生した場合の信号が示される。この場合、ホストプロセッサ211が、部分重みw[4:5]＝01を、１つ上層のメモリチップ層pw4にリアロケートする。この例は、前述の第３のリアロケートパターン1613に対応する。

　ホストプロセッサは、制御信号s1～S3を「001」とする。この制御信号に基づき、マルチプレクサMUX3は、メモリチップ層から読み出した部分重みpw4を元の重みのビット位置にシフトダウンする。更に、ホストプロセッサは、制御信号on1～on4を「0001」とし、マルチプレクサMUX14に固定ビット00を出力させる。残りのマルチプレクサMUX11～MUX13は、読み出された部分重みw[0:1]、w[2:3]と、シフトダウンした部分重みw[4:5]を出力する。この結果、加算器の入力in1は、「01100100」とクロックCK2の場合と同じになる。

　上記２つの例において、加算器の入力in1は、欠陥がない場合「01100110」になり、欠陥がある場合「01100100」になる。欠陥がある場合、最下位ビットの部分重みが「00」となる。よって、プロセッサの演算精度の低下が抑制される。

　以上の通り、第３の実施の形態によれば、複数に分割された部分重みを複数のメモリチップ層にそれぞれ記憶し、ある重みの部分重みを記憶するメモリチップ層のメモリブロックに欠陥が発生した場合、欠陥のある部分重みを、それより下位ビットの部分重みを記憶するメモリチップ層のメモリブロックに記憶する。それにより、欠陥上位ビットの部分重みの使用を継続することができ、NMプロセッサチップの演算精度の低下を抑制できる。

[結語]

　上記の実施の形態では、ニューラルネットワークのパラメータの一つである重みが量子化され、量子化重みが複数の部分重みに分解され、複数の部分重みが複数のメモリチップ層に別々に記憶される。しかし、重み以外のパラメータ、例えば、ニューラルネットワークの畳込み演算のバイアス、等が量子化され、複数の部分パラメータに分解され、複数の部分パラメータが複数のメモリチップ層に別々に記憶されてもよい。

【符号の説明】

201：コンピューティングチップ層

202：ルータ

203：ネットワークインターフェース

204：ネットワークオンチップ

205：ニューラルコンピューティングコア、コア

206：重みメモリ

207：メモリチップ層

208：層間通信部

209：層間ビア

211：ホストプロセッサ

213：NMプロセッサ、ニューロモルフィックプロセッサ

214：チップ間接続

215：パッケージ基板

302：ニューロンモデル

303：重みモデル

305：量子化重み

307：部分重み、部分量子化重み

MB：メモリブロック

MUX：マルチプレクサ

pw1～pw4：部分重み、メモリチップ層の出力

on1-4：制御信号、電源オン・オフ信号、有効・無効信号

s1-s3：制御信号、シフト信号

ADDER：加算器、演算器の一例

in1、in2：加算器の入力

【書類名】特許請求の範囲

【請求項１】

　複数のニューラルコンピューティングコアと、前記複数のニューラルコンピューティングコア間を接続するネットワークオンチップとを有する少なくとも１つのコンピューティングチップ層と、  
　複数のパラメータをそれぞれ記憶する複数のメモリブロックをそれぞれする複数のメモリチップ層とを有し、  
　前記複数のニューラルコンピューティングコアの各々は、前記ネットワークオンチップと接続されるネットワークインターフェースと、ニューロンモデルの演算を実行する複数の物理ニューロンを有するニューロンアレイとを有し、

　前記複数のパラメータは、第１のビット数を有する固定小数点フォーマットに量子化され、  
　前記量子化されたパラメータは、少なくとも、上位ビットを有し前記第１のビット数より小さいビット数を有する第１の部分パラメータと、前記上位ビットより下位の下位ビットを有し前記第１のビット数より小さいビット数を有する第２の部分パラメータに分解され、  
　前記複数のメモリチップ層は、  
　　前記第１の部分パラメータを記憶する第１のメモリチップ層と、  
　　前記第２の部分パラメータを記憶する第２のメモリチップ層を有する、ニューラルネットワークプロセッサ。（第１の実施の形態）

【請求項２】

　前記複数のメモリチップ層は、前記コンピューティングチップ層の上に積層され、前記コンピューティングチップ層及び前記複数のメモリチップ層との間の通信を行う層間ビアが設けられ、

更に、積層された複数のメモリチップ層は、前記第１のメモリチップ層の上に、前記第２のメモリチップ層が積層される、請求項１に記載のニューラルネットワークプロセサ。（第１の実施の形態、積層された複数のメモリチップ層の積層順）

【請求項３】

　前記第１及び第２のメモリチップ層の電源がオン状態になる第１の状態と、前記第１のメモリチップ層の電源がオン状態になり前記第２のメモリチップ層の電源がオフ状態になる第２の状態のいずれかに制御される、請求項２に記載のニューラルネットワークプロセッサ。（第２の実施の形態）

【請求項４】

　前記ニューラルネットワークプロセッサの供給電力が前記ニューラルネットワークプロセッサの消費電力よりも大きい場合、前記第１の状態に制御され、  
　前記供給電力が前記消費電力よりも小さい場合、前記第２の状態に制御される、請求項３に記載のニューラルネットワークプロセッサ。（図９、第２の実施の形態、第１の電力制御の電源オフ制御）

【請求項５】

　前記第２の状態に制御されている状態で、前記供給電力が前記消費電力よりも大きくなると、前記第１の状態に制御される、請求項４に記載のニューラルネットワークプロセッサ。（図１０、第２の実施の形態、第１の電力制御の電源オン制御）

【請求項６】

　前記ニューラルネットワークプロセッサの供給電力が前記ニューラルネットワークプロセッサの消費電力よりも大きい場合、前記第１の状態に制御され、  
　前記第１の状態中に、前記供給電力が前記消費電力よりも小さい場合、前記第２のメモリチップ層の電源電圧を低下するよう制御され、  
　前記第２のメモリチップ層の電源電圧を動作可能な最小電源電圧に低下した状態で、前記供給電力が前記消費電力よりも小さい場合、前記第２の状態に制御される、請求項３に記載のニューラルネットワークプロセッサ。（図１１、第２の実施の形態、第２の電力制御の電源電圧低下と電源オフ制御）

【請求項７】

　前記第２の状態中に、前記供給電力が前記消費電力よりも大きくなると、前記第２のメモリチップ層の電源をオンにし電源電圧を前記最小電源電圧から上昇するよう制御される、請求項６に記載のニューラルネットワークプロセッサ。（図１２、第２の実施の形態、第２の電力制御の電源オンと電源電圧上昇）

【請求項８】

　前記量子化されたパラメータは、最上位ビットを有し前記第１のビット数より小さいビット数を有する第１の部分パラメータから、最下位ビットを有し前記第１のビット数より小さいビット数を有する第Ｎの部分パラメータまで、順に部分パラメータのビットの桁が下がるＮ個の部分パラメータに分解され、前記Ｎは２以上の整数であり、  
　前記積層された複数のメモリチップ層は、  
　前記第１の部分パラメータを記憶する第１のメモリチップ層から、前記第Ｎの部分パラメータを記憶する第Ｎのメモリチップ層まで、順に部分パラメータのビットの桁が下がるＮ個のメモリチップ層として積層され、  
　前記第１から第Ｎのメモリチップ層の電源がオンの状態で、供給電力が消費電力より低くなると、第Ｎから第２のメモリチップ層に向かって各メモリチップ層の電源を順にオフにするよう制御される、請求項３に記載のニューラルネットワークプロセッサ。（第２の実施の形態、図９、図１０）

【請求項９】

　前記第１のメモリチップ層のメモリブロックが欠陥状態になると、前記第１の部分パラメータが前記第２のメモリチップ層に記憶され、前記第２のメモリチップ層から前記第１の部分パラメータが読み出され前記物理ニューロンに入力される、請求項２に記載のニューラルネットワークプロセッサ。（第３の実施の形態）

【請求項１０】

　前記欠陥状態になった第１のメモリチップ層に記憶された前記第１の部分パラメータは読み出しデータから除かれる、請求項９に記載のニューラルネットワークプロセッサ。（第３の実施の形態、図１６）

【請求項１１】

　前記量子化されたパラメータは、最上位ビットを有し前記第１のビット数より小さいビット数を有する第１の部分パラメータから、最下位ビットを有し前記第１のビット数より小さいビット数を有する第Ｎの部分パラメータまで、順に部分パラメータのビットの桁が下がるＮ個の部分パラメータに分解され、前記Ｎは２以上の整数であり、  
　前記積層された複数のメモリチップ層は、  
　前記第１の部分パラメータを記憶する第１のメモリチップ層から、前記第Ｎの部分パラメータを記憶する第Ｎのメモリチップ層まで、最下層から最上層まで順に部分パラメータのビットの桁が下がるＮ個のメモリチップ層として積層され、

前記分解した第１乃至第Ｎの部分パラメータが、前記第１乃至第Ｎのメモリチップ層内の同じアドレスのメモリブロックに記憶され、  
　前記第１乃至第Ｎの部分パラメータのうち第Ｍ（Ｍは１以上Ｎ－１以下の整数）の部分パラメータを記憶するメモリブロックが欠陥状態になると、前記Ｍ乃至第Ｎ－１の部分パラメータが、前記Ｍ＋１乃至第Ｎのメモリチップ層内の前記同じアドレスのメモリブロックにリアロケートして記憶され、前記Ｍ乃至第Ｎ－１の部分パラメータが、前記リアロケートされたメモリブロックから読み出される、請求項９に記載のニューラルネットワークプロセッサ。（第３の実施の形態、図１７、１８）

【請求項１２】

　前記欠陥状態になった第Ｍのメモリチップ層に記憶された前記第Ｍの部分パラメータは読み出しデータから除かれる、請求項１１に記載のニューラルネットワークプロセッサ。（第３の実施の形態、図１７、１８）

【請求項１３】

　前記Ｍを含むリアロケート情報を受信し、

前記物理ニューロン内の演算器は、前記リアロケート情報に基づいて、前記Ｍ＋１乃至第Ｎのメモリチップ層内の前記同じアドレスのメモリブロックから読みだした部分パラメータを、前記Ｍ乃至第Ｎ－１の部分パラメータとして入力する、請求項１１に記載のニューラルネットワークプロセッサ。（第３の実施の形態、図１９、２０）

【請求項１４】

　前記パラメータは、ニューラルネットワークの重み、バイアスのいずれかである、請求項１に記載のニューラルネットワークプロセッサ。

【書類名】要約書

【要約】

【課題】

ニューラルネットワークのパラメータを記憶するメモリ関する課題、コンピューティングコア間の距離、メモリの消費電力、メモリのフォールトトレランス等を解決する。

【解決手段】

複数のコアと、コア間を接続するネットワークオンチップとを有する少なくとも１つのコンピューティングチップ層と、複数のパラメータをそれぞれ記憶する複数のメモリブロックをそれぞれする複数のメモリチップ層とを有し、複数のコアの各々は、前記ネットワークオンチップと接続されるネットワークインターフェースと、ニューロンモデルの演算を実行する複数の物理ニューロンを有するニューロンアレイを有し、前記パラメータは、第１のビット数を有する固定小数点フォーマットに量子化され、前記量子化されたパラメータは、少なくとも、上位ビットを有する第１の部分パラメータと、前記上位ビットより下位の下位ビットを有する第２の部分パラメータに分解され、積層された複数のメモリチップ層は、第１の部分パラメータを記憶する第１のメモリチップ層と、第２の部分パラメータを記憶する第２のメモリチップ層を有するニューラルネットワークプロセッサである。

【選択図】図３