

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7201831号
(P7201831)

(45)発行日 令和5年1月10日(2023.1.10)

(24)登録日 令和4年12月26日(2022.12.26)

(51)国際特許分類	F I
G 1 1 C 16/08 (2006.01)	G 1 1 C 16/08 1 3 0
G 1 1 C 11/54 (2006.01)	G 1 1 C 11/54
G 1 1 C 11/56 (2006.01)	G 1 1 C 11/56 2 1 0

請求項の数 19 (全46頁)

(21)出願番号	特願2021-544169(P2021-544169)	(73)特許権者	500147506
(86)(22)出願日	令和1年9月5日(2019.9.5)		シリコン ストリージ テクノロジー
(65)公表番号	特表2022-523316(P2022-523316 A)		インコーポレイテッド
(43)公表日	令和4年4月22日(2022.4.22)		S I L I C O N S T O R A G E T E C
(86)国際出願番号	PCT/US2019/049817		H N O L O G Y , I N C .
(87)国際公開番号	WO2020/159582		アメリカ合衆国 9 5 1 3 4 カリフォル
(87)国際公開日	令和2年8月6日(2020.8.6)	(74)代理人	ニア州 サンノゼ ホルガー ウェイ 4 5 0
審査請求日	令和3年9月22日(2021.9.22)		110000626
(31)優先権主張番号	62/798,394	(72)発明者	弁理士法人英知国際特許商標事務所
(32)優先日	平成31年1月29日(2019.1.29)		トラン、ヒュー、バン
(33)優先権主張国・地域又は機関	米国(US)	(72)発明者	アメリカ合衆国 9 5 1 3 5 カリフォル
(31)優先権主張番号	16/360,733		ニア州、サンノゼ、ゲイレイ プレイス
(32)優先日	平成31年3月21日(2019.3.21)		2 6 4 2
	最終頁に続く	(72)発明者	ブー、サン
			アメリカ合衆国 9 5 1 3 8 カリフォル
			最終頁に続く

(54)【発明の名称】 深層学習人工ニューラルネットワークにおけるアナログニューラルメモリ用の精密プログラミング回路

(57)【特許請求の範囲】

【請求項1】

不揮発性メモリセルをプログラムするために複数の異なる電圧のうちの1つを生成して、複数の異なる電圧のうちの1つをニューラルネットワーク用の前記不揮発性メモリセルの浮遊ゲートに格納するための調整可能なプログラミング回路であって、

第1の入力端子、第2の入力端子、及び出力端子を備える演算増幅器であって、前記第1の入力端子は、基準電圧を受け取る、演算増幅器と、

前記演算増幅器の前記第2の入力端子に結合された第1の抵抗器と、

前記演算増幅器の前記第2の入力端子及び前記演算増幅器の前記出力端子に結合された第2の抵抗器と、を備え、

前記第1の抵抗器及び前記第2の抵抗器のうちの1つは、可変抵抗器であり、

前記出力端子は、複数の異なる電圧のうちの1つを生成するために、前記可変抵抗器の設定に応答して変動するプログラミング電圧を出力する、調整可能なプログラミング回路。

【請求項2】

前記第1の抵抗器は、前記可変抵抗器である、請求項1に記載の調整可能なプログラミング回路。

【請求項3】

前記第2の抵抗器は、前記可変抵抗器である、請求項1に記載の調整可能なプログラミング回路。

【請求項4】

前記第 1 の抵抗器は、低電圧ドメインにおいて動作する前記可変抵抗器である、請求項 1 に記載の調整可能なプログラミング回路。

【請求項 5】

前記第 2 の抵抗器は、高電圧ドメインにおいて動作する前記可変抵抗器である、請求項 1 に記載の調整可能なプログラミング回路。

【請求項 6】

前記演算増幅器用の高電圧電源は、前記演算増幅器の出力電圧に基づいて適応可能である、請求項 1 に記載の調整可能なプログラミング回路。

【請求項 7】

前記不揮発性メモリセルは、積層ゲートメモリセルである、請求項 1 に記載の調整可能なプログラミング回路。

10

【請求項 8】

前記不揮発性メモリセルは、スプリットゲートメモリセルである、請求項 1 に記載の調整可能なプログラミング回路。

【請求項 9】

不揮発性メモリセルをプログラムするために複数の異なる電圧のうちの 1 つを生成して、複数の異なる電圧のうちの 1 つをニューラルネットワーク用の前記不揮発性メモリセルの浮遊ゲートに格納するための調整可能なプログラミング回路であって、

第 1 の入力端子、第 2 の入力端子、及び出力端子を備える第 1 の演算増幅器であって、前記第 1 の入力端子は、基準電圧を受け取る、第 1 の演算増幅器と、

20

第 1 の入力端子、第 2 の入力端子、及び出力端子を備える第 2 の演算増幅器であって、前記第 2 の演算増幅器の前記第 1 の入力端子は、前記第 2 の演算増幅器の前記出力端子に結合された、第 2 の演算増幅器と、

前記第 1 の演算増幅器の前記第 2 の入力端子及び前記第 1 の演算増幅器の前記出力端子に結合された第 1 の抵抗器と、

前記第 1 の演算増幅器の前記第 2 の入力端子と前記第 2 の演算増幅器の前記出力端子との間に結合された第 2 の抵抗器と、

前記第 2 の演算増幅器の前記第 2 の入力端子に結合された調整可能な分圧器と、を備え、

前記第 1 の演算増幅器の前記出力端子は、前記調整可能な分圧器によって出力された電圧に応答して変動するプログラミング電圧を出力する、調整可能なプログラミング回路。

30

【請求項 10】

前記調整可能な分圧器及び前記第 2 の演算増幅器は、低電圧ドメインにおいて動作する、請求項 9 に記載の調整可能なプログラミング回路。

【請求項 11】

前記第 1 の演算増幅器は、高電圧ドメインにおいて動作する、請求項 9 に記載の調整可能なプログラミング回路。

【請求項 12】

前記第 1 の演算増幅器用の前記高電圧電源は、前記プログラミング電圧に基づいて調整可能である、請求項 9 に記載の調整可能なプログラミング回路。

【請求項 13】

40

前記不揮発性メモリセルは、積層ゲートメモリセルである、請求項 9 に記載の調整可能なプログラミング回路。

【請求項 14】

前記不揮発性メモリセルは、スプリットゲートメモリセルである、請求項 9 に記載の調整可能なプログラミング回路。

【請求項 15】

不揮発性メモリセルをプログラムするために複数の異なる電圧のうちの 1 つを生成して、複数の異なる電圧のうちの 1 つをニューラルネットワーク用の前記不揮発性メモリセルの浮遊ゲートに格納するための調整可能なプログラミング回路であって、

第 1 の出力端子及び第 2 の出力端子を備える第 1 の調整可能な分圧器と、

50

第2の調整可能な分圧器と、

第1の入力端子、第2の入力端子、及び出力端子を備える第1の演算増幅器であって、前記第1の演算増幅器の前記第1の入力端子は、前記第1の演算増幅器の前記出力端子に結合され、前記第1の演算増幅器の前記第2の入力端子は、前記第1の調整可能な分圧器の前記第1の出力端子に結合され、前記第1の演算増幅器の前記出力端子は、第1のプログラミング電圧を提供する、第1の演算増幅器と、

第1の入力端子、第2の入力端子、及び出力端子を備える第2の演算増幅器であって、前記第2の演算増幅器の前記第1の入力端子は、前記第2の演算増幅器の前記出力端子に結合され、前記第2の演算増幅器の前記第1の入力端子は、前記第1の調整可能な分圧器の前記第2の出力端子に結合された、第2の演算増幅器と、を備え、

10

前記第2の調整可能な分圧器は、前記第2の演算増幅器の前記出力端子及び前記第1の演算増幅器の前記出力端子に結合され、第2のプログラミング電圧を提供し、

前記第1のプログラミング電圧は、前記第1の調整可能な分圧器に応答して変動し、前記第2のプログラミング電圧は、前記第1の調整可能な分圧器及び前記第2の調整可能な分圧器に応答して変動する、調整可能なプログラミング回路。

【請求項16】

前記第1の分圧器は、粗レベル調整能力を提供し、前記第2の分圧器は、微細レベル調整能力を提供する、請求項15に記載の調整可能なプログラミング回路。

【請求項17】

前記演算増幅器用の前記高電圧電源は、前記プログラミング電圧に基づいて調整可能である、請求項15に記載の調整可能なプログラミング回路。

20

【請求項18】

前記不揮発性メモリセルは、積層ゲートメモリセルである、請求項15に記載の調整可能なプログラミング回路。

【請求項19】

前記不揮発性メモリセルは、スプリットゲートメモリセルである、請求項15に記載の調整可能なプログラミング回路。

【発明の詳細な説明】

【技術分野】

【0001】

30

(優先権の主張)

本出願は、2019年1月29日に出願された「Precision Programming Circuit For Analog Neural Memory In Deep Learning Artificial Neural Network」と題する米国仮特許出願第62/798,394号、及び2019年3月21日に出願された「Precision Programming Circuit For Analog Neural Memory In Deep Learning Artificial Neural Network」と題する米国特許出願第16/360,733号の優先権を主張する。

【0002】

40

(発明の分野)

多くの可能な値のうちの1つを格納するために、アナログニューラルメモリ内の不揮発性メモリセルをプログラミングするための多数の精密プログラミング回路が開示される。

【背景技術】

【0003】

人工ニューラルネットワークは、生物学的ニューラルネットワーク（動物の中樞神経系、特に脳）を模倣しており、多数の入力に依存し得、かつ、一般的に未知である関数を推定する又は近似するために使用される。人工ニューラルネットワークは、概して、メッセージを交換する相互接続した「ニューロン」の層を含む。

【0004】

50

図1は人工ニューラルネットワークを示しており、図中、円は、入力又はニューロンの層を表す。接続（シナプスと呼ばれる）は、矢印によって表され、経験に基づいて調整され得る数値の重みを有する。これにより、ニューラルネットワークは入力に適応し、学習可能になる。典型的には、ニューラルネットワークは、複数の入力の層を含む。典型的には、1つ以上のニューロンの中間層、及びニューラルネットワークの出力を提供するニューロンの出力層が存在する。各レベルにおけるニューロンは、シナプスから受信したデータに基づいて個々に又は集合的に決定を行う。

【0005】

高性能情報処理用の人工ニューラルネットワークの開発における主要な課題の1つは、適切なハードウェア技術の欠如である。実際には、実用ニューラルネットワークは、非常に多数のシナプスに依拠しており、これによりニューロン間の高い接続性、すなわち、非常に高度な計算処理の並列化が可能となる。原理的には、このような複雑性は、デジタルスーパーコンピュータ又は専用GPU（グラフィックプロセッシングユニット）クラスタによって実現が可能である。しかしながら、高コストに加え、これらのアプローチはまた、生物学的ネットワークが主として低精度のアナログ計算を実行するのではるかに少ないエネルギーしか消費しないのと比較して、エネルギー効率が劣っていることに悩まされている。人工ニューラルネットワークにはCMOSアナログ回路が使用されてきたが、ほとんどのCMOS実装シナプスは、必要な多数のニューロン及びシナプスを前提とすると、嵩高過ぎていた。

【0006】

出願人は以前に、参照により組み込まれる、米国特許公開第2017/0337466号として公開された米国特許出願第15/594,439号において、シナプスとして1つ以上の不揮発性メモリアレイを利用する人工（アナログ）ニューラルネットワークを開示した。不揮発性メモリアレイは、アナログニューロモーフィックメモリとして動作する。ニューラルネットワークデバイスは、第1の複数の入力を受信して、それから第1の複数の出力を生成するように構成された第1の複数のシナプス、及び第1の複数の出力を受信するように構成された第1の複数のニューロンを含む。第1の複数のシナプスは、複数のメモリセルを含み、各メモリセルは、半導体基板内に形成され、間にチャネル領域が延在する離間したソース領域及びドレイン領域と、チャネル領域の第1の部分の上方に配設され、チャネル領域の第1の部分から絶縁された浮遊ゲートと、チャネル領域の第2の部分の上方に配設され、チャネル領域の第2の部分から絶縁された非浮遊ゲートと、を含む。複数のメモリセルの各々は、浮遊ゲート上の多くの電子に対応する重み値を格納するように構成されている。複数のメモリセルは、第1の複数の入力に、格納された重み値を乗算して第1の複数の出力を生成するように構成される。

【0007】

アナログニューロモーフィックメモリスシステムに使用される各不揮発性メモリセルは、消去・プログラムに対応して、浮遊ゲート内に電荷、すなわち電子の数、を非常に特異的かつ正確な量で保持しなければならない。例えば、各浮遊ゲートはN個の異なる値のうちの1つを保持しなければならない。ここで、Nは、各セルによって示され得る異なる重みの数である。Nの例としては、16、32、64、128及び256が挙げられる。

【0008】

VMMシステムにおける1つの課題は、非常に特異的かつ正確な量の電荷を保持するように選択されたセルの浮遊ゲートをプログラムし、それにより、各浮遊ゲートが異なるN個の値のうちの1つを保持することができるようにする必要があるため、VMMのメモリセルを正確にプログラムする能力である。

【0009】

必要とされるのは、VMM内のメモリセルを正確にプログラミングするための改善されたシステムである。

【発明の概要】

【0010】

10

20

30

40

50

多くの可能な値のうちの1つを格納するために、アナログニューラルメモリ内のVMM内の不揮発性メモリセルをプログラミングするための多数の精密プログラミング回路が開示される。

【0011】
【0012】
【0013】
【0014】
【0015】
【0016】
【0017】
【0018】
【0019】
【0020】
【0021】
【0022】
【0023】
【0024】
【0025】
【0026】
【0027】
【0028】
【0029】
【0030】
【0031】
【0032】
【0033】
【0034】
【0035】
【0036】
【0037】
【0038】
【0039】
【0040】
【0041】
【0042】
【0043】
【0044】
【0045】
【0046】
【0047】
【0048】
【0049】
【0050】
【0051】
【0052】
【0053】
【0054】
【0055】
【0056】
【0057】

10

20

30

40

50

【0058】

【0059】

【0060】

【0061】

【0062】

【0063】

【0064】

【0065】

【0066】

【図面の簡単な説明】

10

【0067】

【図1】 先行技術の人工ニューラルネットワークを示す図である。

【図2】 先行技術のスプリットゲート型フラッシュメモリセルを示す。

【図3】 別の先行技術のスプリットゲート型フラッシュメモリセルを示す。

【図4】 別の先行技術のスプリットゲート型フラッシュメモリセルを示す。

【図5】 別の先行技術のスプリットゲート型フラッシュメモリセルを示す。

【図6】 別の先行技術のスプリットゲート型フラッシュメモリセルを示す。

【図7】 先行技術の積層ゲートフラッシュメモリセルを示す。

【図8】 1つ以上の不揮発性メモリアレイを利用する例示的な人工ニューラルネットワークの異なるレベルを示す図である。

20

【図9】 ベクトル行列乗算システムを示すブロック図である。

【図10】 1つ以上のベクトル行列乗算システムを利用する例示的な人工ニューラルネットワークを示すブロック図である。

【図11】 ベクトル行列乗算システムの別の実施形態を示す。

【図12】 ベクトル行列乗算システムの別の実施形態を示す。

【図13】 ベクトル行列乗算システムの別の実施形態を示す。

【図14】 ベクトル行列乗算システムの別の実施形態を示す。

【図15】 ベクトル行列乗算システムの別の実施形態を示す。

【図16】 先行技術の長・短期記憶システムを示す。

【図17】 長・短期記憶システムで使用する例示的なセルを示す。

30

【図18】 図17の例示的なセルの一実施形態を示す。

【図19】 図17の例示的なセルの別の実施形態を示す。

【図20】 先行技術のゲート付き回帰型ユニットシステムを示す。

【図21】 ゲート付き回帰型ユニットシステムで使用する例示的なセルを示す。

【図22】 図21の例示的なセルの一実施形態を示す。

【図23】 図21の例示的なセルの別の実施形態を示す。

【図24】 ベクトル行列乗算システムの別の実施形態を示す。

【図25】 ベクトル行列乗算システムの別の実施形態を示す。

【図26】 ベクトル行列乗算システムの別の実施形態を示す。

【図27】 ベクトル行列乗算システムの別の実施形態を示す。

40

【図28】 ベクトル行列乗算システムの別の実施形態を示す。

【図29】 ベクトル行列乗算システムの別の実施形態を示す。

【図30】 ベクトル行列乗算システムの別の実施形態を示す。

【図31】 ベクトル行列乗算システムの別の実施形態を示す。

【図32】 VMMシステムを示す。

【図33】 アナログニューロメモリシステムを示す。

【図34】 ベクトル乗算器マトリックスシステムと共に使用するための高電圧生成ブロックを示す。

【図35】 チャージポンプ及びチャージポンプ調整回路を示す。

【図36】 電流補償回路を有する高電圧生成ブロックを示す。

50

【図 3 7】電流補償回路を有する別の高電圧生成ブロックを示す。

【図 3 8】別の高電圧生成ブロックを示す。

【図 3 9】電流補償を提供するためのダミービット線を示す。

【図 4 0】高電圧デコーダを示す。

【図 4 1】高電圧試験回路を示す。

【図 4 2】高電圧生成ブロックを示す。

【図 4 3】別の高電圧生成ブロックを示す。

【図 4 4】別の高電圧生成ブロックを示す。

【図 4 5】高電圧演算増幅器を示す。

【図 4 6】別の高電圧演算増幅器を示す。

10

【図 4 7】適応高電圧電源を示す。

【図 4 8】列ドライバを示す。

【図 4 9】列センス増幅器を示す。

【図 5 0】読み出し基準回路を示す。

【図 5 1】別の読み出し基準回路を示す。

【図 5 2】適応高電圧電源を示す。

【図 5 3】別の適応高電圧電源を示す。

【図 5 4】別の適応高電圧電源を示す。

【図 5 5】別の適応高電圧電源を示す。

【図 5 6】別の適応高電圧電源を示す。

20

【図 5 7】単一基準検証アルゴリズムを示す。

【図 5 8】二重基準検証アルゴリズムを示す。

【図 5 9】調整可能な基準電圧源を示す。

【図 6 0】図 5 9 の調整可能な基準電圧源で使用されるサブ回路を示す。

【発明を実施するための形態】

【0068】

本発明の人工ニューラルネットワークは、CMOS 技術及び不揮発性メモリアレイの組み合わせを利用する。

不揮発性メモリセル

【0069】

30

デジタル不揮発性メモリは、周知である。例えば、参照により本明細書に組み込まれる、米国特許第 5,029,130 号（「130 号特許」）は、フラッシュメモリセルの一種である、スプリットゲート型不揮発性メモリセルのレイアウトを開示する。このようなメモリセル 210 を図 2 に示す。各メモリセル 210 は、半導体基板 12 内に形成されたソース領域 14 とドレイン領域 16 と、を含み、ソース領域 14 とドレイン領域 16 の間にはチャネル領域 18 がある。浮遊ゲート 20 は、チャネル領域 18 の第 1 の部分の上方に形成され、チャネル領域 18 の第 1 の部分から絶縁され（並びに、チャネル領域 18 の第 1 の部分の導電性を制御して）、ソース領域 14 の一部分の上方にかけて形成される。ワード線端子 22（典型的には、ワード線に結合される）は、チャネル領域 18 の第 2 の部分の上方に配設され、チャネル領域 18 の第 2 の部分から絶縁された、（並びに、チャネル領域 18 の第 2 の部分の導電性を制御する）第 1 の部分と、上に向かって浮遊ゲート 20 の上方にかけて延在する第 2 の部分と、を有する。浮遊ゲート 20 及びワード線端子 22 は、ゲート酸化物によって基板 12 から絶縁される。ビット線 24 はドレイン領域 16 に結合される。

40

【0070】

ワード線端子 22 に高圧正電圧を印加することによって、メモリセル 210 に対して消去が行われ（電子が浮遊ゲートから除去される）、これによって、浮遊ゲート 20 の電子は、浮遊ゲート 20 からワード線端子 22 までそれらの間にある絶縁体の中をファウラーノルドハイム（Fowler-Nordheim）トンネリングを介して通過する。

【0071】

50

メモリセル 210 は、ワード線端子 22 に正電圧、及びソース領域 14 に正電圧を印加することによってプログラムされる（電子が浮遊ゲートに印加される）。電子電流は、ソース領域 14 からドレイン領域 16 に向かって流れる。電子は加速し、ワード線端子 22 と浮遊ゲート 20 との間の間隙に達すると、発熱する。熱せられた電子の一部が、浮遊ゲート 20 からの静電引力に起因して、浮遊ゲート 20 にゲート酸化物を介して注入される。
【0072】

メモリセル 210 は、ドレイン領域 16 及びワード線端子 22 に正の読み出し電圧を印加する（ワード線端子の下方のチャネル領域 18 の部分をオンにする）ことによって、読み出される。浮遊ゲート 20 が正に帯電する（すなわち、電子を消去する）と、浮遊ゲート 20 の下方のチャネル領域 18 の部分も同様にオンになり、電流はチャネル領域 18 を流れ、これは、消去された状態つまり「1」の状態として検知される。浮遊ゲート 20 が負に帯電する（すなわち、電子でプログラムされる）と、浮遊ゲート 20 の下方のチャネル領域の部分はほとんど又は完全にオフになり、電流はチャネル領域 18 を流れず（又はほとんど流れず）、これは、プログラムされた状態つまり「0」の状態として検出される。
【0073】

表 1 は、読み出し、消去、及びプログラム動作を実行するためのメモリセル 110 の端子に印加され得る典型的な電圧範囲を示す。

表 1：図 2 のフラッシュメモリセル 210 の動作

【表 1】

	WL	BL	SL
読み出し	2~3V	0.6~2V	0V
消去	約11~13V	0V	0V
プログラム	1~2V	1~3 μ A	9~10V

【0074】

図 3 は、制御ゲート（control gate、CG）28を追加した、図 2 のメモリセル 210 に類似しているメモリセル 310 を示す。制御ゲート 28 は、プログラミングでは、例えば 10V の高電圧で、消去では、例えば 0V / -8V の低又は負で、読み出しでは、例えば 0V / 2.5V の低又は中範囲でバイアスされる。他の端子は、図 2 の端子と同様にバイアスされる。

【0075】

図 4 は、ソース領域 14 と、ドレイン領域 16 と、チャネル領域 18 の第 1 の部分の上方にある浮遊ゲート 20 と、チャネル領域 18 の第 2 の部分の上方にある選択ゲート 22（典型的には、ワード線、WL、に結合される）と、浮遊ゲート 20 の上方にある制御ゲート 28 と、ソース領域 14 の上方にある消去ゲート 30 と、を備える 4 ゲートメモリセル 410 を示す。この構成は、あらゆる目的のため参照により本明細書に組み込まれる、米国特許第 6,747,310 号に記載されている。ここで、全てのゲートは、浮遊ゲート 20 を除いて、非浮遊ゲートであり、つまり、それらは電圧源に電氣的に接続される又は接続可能である。プログラミングは、熱せられた電子がチャネル領域 18 から浮遊ゲート 20 にその電子自体を注入することによって実行される。消去は、電子が浮遊ゲート 20 から消去ゲート 30 へトンネリングすることによって実行される。

【0076】

表 2 は、読み出し、消去、及びプログラム動作を実行するためのメモリセル 310 の端子に印加され得る典型的な電圧範囲を示す。

表 2：図 4 のフラッシュメモリセル 410 の動作

10

20

30

40

50

【表 2】

	WL/SG	BL	CG	EG	SL
読み出し	1.0~2V	0.6~2V	0~2.6V	0~2.6V	0V
消去	-0.5V/0V	0V	0V/-8V	8~12V	0V
プログラム	1V	1 μ A	8~11V	4.5~9V	4.5~5V

【0077】

図5は、メモリセル510が消去ゲートEGを含まないことを除いて、図4のメモリセル410に類似しているメモリセル510を示す。消去は、基板18を高電圧にバイアスし、制御ゲートCG28を低電圧又は負電圧にバイアスすることによって実行される。代替的に、消去は、ワード線22を正電圧にバイアスし、制御ゲート28を負電圧にバイアスすることによって実行される。プログラミング及び読み出しは、図4のものに類似している。

【0078】

図6は、別の種類のフラッシュメモリセルである、3ゲートメモリセル610を示す。メモリセル610は、メモリセル610が別個の制御ゲートを有しないことを除いて、図4のメモリセル410と同一である。消去動作（消去ゲートの使用を通して消去が起こる）及び読み出し動作は、制御ゲートバイアスが印加されないことを除いて、図4のものに類似している。プログラミング動作もまた、制御ゲートバイアスなしで行われるため、結果として、プログラム動作中は、制御ゲートバイアスの不足を補償するため、より高い電圧がソース線に印加されなければならない。

【0079】

表3は、読み出し、消去、及びプログラム動作を実行するためのメモリセル610の端子に印加され得る典型的な電圧範囲を示す。

表3：図6のフラッシュメモリセル610の動作

【表3】

	WL/SG	BL	EG	SL
読み出し	0.7~2.2V	0.6~2V	0~2.6V	0V
消去	-0.5V/0V	0V	11.5V	0V
プログラム	1V	2~3 μ A	4.5V	7~9V

【0080】

図7は、別の種類のフラッシュメモリセルである、積層ゲートメモリセル710を示す。メモリセル710は、浮遊ゲート20がチャネル領域18全体の上方にわたって延在し、制御ゲート22（ワード線に結合される）が絶縁層（図示せず）によって分離された浮遊ゲート20の上方に延在することを除いて、図2のメモリセル210に類似している。消去、プログラミング、及び読み出し動作は、メモリセル210について前述したものと同様の様態において動作する。

【0081】

表4は、読み出し、消去、及びプログラム動作を実行するためのメモリセル710及び基板12の端子に印加され得る典型的な電圧範囲を示す。

表4：図7のフラッシュメモリセル710の動作

10

20

30

40

50

【表 4】

	CG	BL	SL	基板
読み出し	2～5V	0.6～2V	0V	0V
消去	-8～-10V/0V	FLT	FLT	8～10V/15～20V
プログラム	8～12V	3～5V	0V	0V

【0082】

上記の人工ニューラルネットワークにおける不揮発性メモリセルの種類のうちの1つを備えるメモリアレイを利用するために、2つの修正が行われる。第1に、以下に更に説明されるように、アレイ内の他のメモリセルのメモリ状態に悪影響を与えずに各メモリセルを個々にプログラム、消去、及び読み出しできるように線を構成する。第2に、メモリセルの連続（アナログ）プログラミングを提供する。

【0083】

具体的には、アレイ内の各メモリセルのメモリ状態（すなわち、浮遊ゲート上の電荷。）を、完全に消去された状態から完全にプログラムされた状態へ、独立して、かつ他のメモリセルの異常が最小で、連続的に変えることができる。別の実施形態では、アレイ内の各メモリセルのメモリ状態（すなわち、浮遊ゲートの電荷）を、完全にプログラムされた状態から完全に消去された状態へ、及び逆もまた同様に、独立して、かつ他のメモリセルの異常が最小で、連続的に変えることができる。これはつまり、セル格納がアナログであるか、又は多数の離散値（16又は64の異なる値など）のうちの1つを最低限格納できることを意味し、これにより、メモリアレイ内の全てのセルが非常に正確に、かつ個別に調整可能となり、またメモリアレイは格納に対して理想的になり、ニューラルネットワークのシナプシスの重みに微調整が可能となる。

【0084】

本明細書に記載される方法及び手段は、限定されることなく、SONOS（ケイ素-酸化物-窒化物-酸化物-ケイ素、窒化物中の電荷トラップ）、MONOS（金属-酸化物-窒化物-酸化物-ケイ素、窒化物中の金属電荷トラップ）、ReRAM（抵抗ram）、PCM（相変化メモリ）、MRAM（磁気ram）、FeRAM（強誘電性ram）、OTP（バイレベル又はマルチレベルのワンタイムプログラマブル）、及びCeRAM（相関電子ram）など、他の不揮発性メモリ技術に適用することができる。本明細書に記載される方法及び手段は、限定されることなく、SRAM、DRAM、及び揮発性シナプスセルなど、ニューラルネットワークに使用される揮発性メモリ技術に適用することができる。

不揮発性メモリセルアレイを使用するニューラルネットワーク

【0085】

図8は、本実施形態の不揮発性メモリアレイを利用するニューラルネットワークの非限定実施例を概念的に示す。この例は、顔認識アプリケーション用に不揮発性メモリアレイニューラルネットワークを使用するが、不揮発性メモリアレイベースのニューラルネットワークを使用して他の適切なアプリケーションを実装することも可能である。

【0086】

S0は入力層であり、この例では、5ビット精度の32×32ピクセルRGB画像である（すなわち、各色R、G、及びBにつき1つずつで3つの32×32ピクセルアレイであり、各ピクセルは5ビット精度である）。入力層S0から層C1に行くシナプスCB1は、一部のインスタンスには異なる重みのセットを適用し、他のインスタンスには共有の重みを適用し、入力画像を3×3ピクセルの重なり合うフィルタでスキャンし（カーネル）、1ピクセル（又はモデルによっては2ピクセル以上）ずつフィルタをシフトする。具体的には、画像の3×3部分における9ピクセルの値（すなわち、フィルタ又はカーネルと称される）は、シナプスCB1に提供され、そこでこれらの9個の入力値に適切な重み

を乗算し、その乗算の出力を合計後、単一の出力値が決定され、特徴マップ C 1 の層の 1 つのピクセルを生成するために C B 1 の第 1 のシナプスによって与えられる。3 × 3 フィルタは次に、入力層 S 0 内で右側に 1 ピクセルだけシフトされ（すなわち、3 ピクセルの列を右側に追加し、左側で 3 ピクセルの列をドロップする）、これにより、この新しく位置づけられたフィルタの 9 ピクセル値は、シナプス C B 1 に提供され、そこで、それらに上記と同じ重みを乗算し、関連するシナプスによって第 2 の単一の出力値を決定する。このプロセスを、3 × 3 フィルタが入力層 S 0 の 3 2 × 3 2 ピクセル画像全体にわたって 3 色全て及び全てのビット（精度値）についてスキャンするまで続ける。プロセスは次に、層 C 1 の特徴マップ全てが計算されるまで、異なる重みのセットを使用して繰り返されて、C 1 の異なる特徴マップを生成する。

10

【0087】

本例では、層 C 1 において、各々 3 0 × 3 0 ピクセルを有する 1 6 個の特徴マップが存在する。各ピクセルは、入力とカーネルとの乗算から抽出された新しい特徴ピクセルであり、したがって、各特徴マップは、2 次元アレイであり、したがってこの例では、層 C 1 は、2 次元アレイの 1 6 層を構成する（本明細書で言及される層及びアレイは、必ずしも物理的関係ではなく論理的な関係であり、すなわち、アレイは必ずしも物理的な 2 次元アレイに配向されないことに留意されたい）。層 C 1 内の 1 6 個の特徴マップの各々は、フィルタスキャンに適用される、1 6 個の異なるシナプス重みのセットのうちの 1 つによって生成される。C 1 特徴マップは全て、境界識別など、同じ画像特徴の異なる態様を対象とすることができる。例えば、第 1 のマップ（この第 1 のマップを生成するために使用される全てのスキャンに共有される第 1 の重みセットを使用して生成される）は、円形エッジを識別することができ、第 2 のマップ（第 1 の重みセットとは異なる第 2 の重みセットを使用して生成される）は、方形エッジ又は特定の特徴のアスペクト比などを識別することができる。

20

【0088】

層 C 1 から層 S 1 へ行く前には、各特徴マップ内の重なり合わずに連続する 2 × 2 領域からの値をプールの活性化関数 P 1（プーリング）が適用される。プーリング関数の目的は、近隣の位置を平均すること（又は max 関数を使用することも可能である）、例えばエッジ位置の依存を低減すること、及び次の段階に行く前にデータサイズを低減することである。層 S 1 において、1 6 個の 1 5 × 1 5 特徴マップ（すなわち、各々 1 5 × 1 5 ピクセルの 1 6 個の異なるアレイ）が存在する。層 S 1 から層 C 2 に行くシナプス C B 2 は、S 1 内のマップを 4 × 4 フィルタにより 1 ピクセルのフィルタシフトでスキャンする。層 C 2 において、2 2 個の 1 2 × 1 2 特徴マップが存在する。層 C 2 から層 S 2 へ行く前には、各特徴マップ内の重なり合わずに連続する 2 × 2 領域からの値をプールの活性化関数 P 2（プーリング）が適用される。層 S 2 において、2 2 個の 6 × 6 特徴マップが存在する。層 S 2 から層 C 3 へ行くシナプス C B 3 では活性化関数（プーリング）が適用され、ここで層 C 3 内の全てのニューロンは、C B 3 のそれぞれのシナプスを介して層 S 2 内の全てのマップに接続する。層 C 3 において、6 4 個のニューロンが存在する。層 C 3 から出力層 S 3 へと行くシナプス C B 4 は、C 3 を S 3 に完全に接続する、すなわち、層 C 3 内の全てのニューロンは、層 S 3 内の全てのニューロンに接続される。S 3 における出力は、1 0 個のニューロンを含み、ここで出力が最も高いニューロンが、クラスを決定する。この出力は、例えば、元の画像の内容の識別又は分類を示すことができる。

30

40

【0089】

シナプスの各層は、不揮発性メモリセルのアレイ又はアレイの一部を使用して実装される。

【0090】

図 9 は、その目的のために使用可能なアレイのブロック図である。ベクトル行列乗算（Vector-by-matrix multiplication、VMM）アレイ 3 2 は、不揮発性メモリセルを含み、ある層と次の層との間のシナプス（図 6 の C B 1、C B 2、C B 3、及び C B 4 など）として利用される。具体的には、VMM アレイ 3 2 は、不揮発性メモリセルのアレイ 3 3

50

、消去ゲート及びワード線ゲートデコーダ34、制御ゲートデコーダ35、ビット線デコーダ36、並びにソース線デコーダ37を含み、それらのデコーダは不揮発性メモリセルアレイ33に対するそれぞれの入力をデコードする。VMMアレイ32への入力は、消去ゲート及びワード線ゲートデコーダ34から、又は制御ゲートデコーダ35から行うことができる。この例におけるソース線デコーダ37はまた、不揮発性メモリセルアレイ33の出力をデコードする。代替的に、ビット線デコーダ36が、不揮発性メモリセルアレイ33の出力をデコードすることができる。

【0091】

不揮発性メモリセルアレイ33は、2つの目的を果たす。第1に、VMMアレイ32によって使用される重みを格納する。第2に、不揮発性メモリセルアレイ33は、不揮発性メモリセルアレイ33に格納された重みによって、入力を有効に乗算して、それらを出線（ソース線又はビット線）ごとに加算して、出力を生成し、この出力は次の層への入力又は最後の層への入力になる。不揮発性メモリセルアレイ33が乗算及び加算の関数を実行することで、別個の乗算及び加算の論理回路の必要性はなくなり、また、その場でのメモリ内の計算により電力効率も良い。

【0092】

不揮発性メモリセルアレイ33の出力は、不揮発性メモリセルアレイ33の出力を合計してその畳み込み用の単一の値を作成する、差動加算器（加算オペアンプ又は加算電流ミラーなど）38に供給される。差動加算器38は、正の重み及び負の重みの総和を実行するように配置される。

【0093】

差動加算器38の合計された出力値は、次に出力を整流する活性化関数回路39に供給される。活性化関数回路39は、シグモイド、 $\tanh$ 、ReLU関数、又は任意の他の非線形関数を提供し得る。活性化関数回路39の整流された出力値は、次の層（例えば図8のC1）の特徴マップの要素になり、次いで、次のシナプスに適用されて次の特徴マップ層又は最後の層を生成する。したがって、この例では、不揮発性メモリセルアレイ33は、複数のシナプスを構成し（ニューロンの前の層から、又は画像データベースなどの入力層から、入力を受信する）、加算器38及び活性化関数回路39は、複数のニューロンを構成する。

【0094】

図9のVMMアレイ32への入力（WLx、EGx、CGx、及び任意選択的にBLx及びSLx）は、アナログレベル、バイナリレベル、デジタルパルス（この場合、パルスを適切な入力アナログレベルに変換するためにパルスアナログ変換器PACが必要となり得る）、又はデジタルビット（この場合、デジタルビットを適切な入力アナログレベルに変換するためにDACが提供される）であり得、出力は、アナログレベル、バイナリレベル、デジタルパルス、又はデジタルビット（この場合、出力アナログレベルをデジタルビットに変換するために出力ADCが提供される）であり得る。

【0095】

図10は、図中でVMMアレイ32a、32b、32c、32d及び32eとして標示されたVMMアレイ32の多数の層の使用を示すブロック図である。図10に示されるように、入力（Inputxで示される）は、デジタルアナログ変換器31によってデジタルからアナログに変換され、入力VMMアレイ32aに提供される。変換されたアナログ入力は、電圧又は電流であり得る。第1の層の入力D/A変換は、入力VMMアレイ32aの行列乗算器の適切なアナログレベルに入力Inputxをマッピングする関数又はLUT（ルックアップテーブル）を使用することによって行うことができる。入力変換はまた、外部アナログ入力を入力VMMアレイ32aへのマッピングされたアナログ入力に変換するために、アナログーアナログ（A/A）変換器によって行うこともできる。入力変換はまた、外部デジタル入力を入力VMMアレイ32aへのマッピングされた1つ又は複数のデジタルパルスに変換するために、デジタルーデジタルパルス（D/P）変換器によって行うこともできる。

10

20

30

40

50

【0096】

入力VMMアレイ32aによって生成された出力は、次に、次のVMMアレイ（隠しレベル1）32bへの入力として提供され、次に入力VMMアレイ（隠しレベル2）32cへの入力として提供される出力を生成する、などとなる。VMMアレイ32の様々な層は、畳み込みニューラルネットワーク（convolutional neural network、CNN）のシナプス及びニューロンの各層として機能する。各VMMアレイ32a、32b、32c、32d及び32eは、スタンドアローンの物理的不揮発性メモリアレイとすることができ、又は複数のVMMアレイは、同じ物理的不揮発性メモリアレイの異なる部分を利用することができ、又は複数のVMMアレイは、同じ物理的不揮発性メモリアレイの重なり合う部分を利用することができる。各VMMアレイ32a、32b、32c、32d、及び32eはまた、そのアレイ又はニューロンの様々な部分に対して時分割多重化され得る。図10に示される例は、5つの層（32a、32b、32c、32d、32e）、すなわち、1つの入力層（32a）、2つの隠れ層（32b、32c）、及び2つの完全に接続された層（32d、32e）を含む。当業者であれば、これは単なる例示であり、代わりにシステムが2つを超える隠れ層及び2つを超える完全に接続された層を備え得ることを理解するであろう。

10

ベクトル行列乗算（VMM）アレイ

【0097】

図11は、図3に示されるメモリセル310に特に適したニューロンVMMアレイ1100を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。VMMアレイ1100は、不揮発性メモリセルのメモリアレイ1101と、不揮発性基準メモリセルの基準アレイ1102（アレイの頂部に位置する）と、を備える。代替的に、別の基準アレイが底部に位置することができる。

20

【0098】

VMMアレイ1100では、制御ゲート線1103などの制御ゲート線が垂直方向に延びており（したがって、行方向の基準アレイ1102が、制御ゲート線1103に直交する）、消去ゲート線1104などの消去ゲート線が水平方向に延びている。ここで、VMMアレイ1100への入力、制御ゲート線（CG0、CG1、CG2、CG3）に提供され、VMMアレイ1100の出力はソース線（SL0、SL1）に現れる。一実施形態では、偶数行のみが使用され、別の実施形態では、奇数行のみが使用される。各ソース線（それぞれSL0、SL1）に加えられる電流は、その特定のソース線に接続されたメモリセルからの全ての電流の合計関数を実行する。

30

【0099】

ニューラルネットワークについて本明細書に記載されているように、VMMアレイ1100の不揮発性メモリセル、すなわちVMMアレイ1100のフラッシュメモリは、好ましくは、サブスレッショルド領域において動作するように構成される。

【0100】

本明細書に記載される不揮発性基準メモリセル及び不揮発性メモリセルは、以下のように弱い反転でバイアスされる：

$$I_{ds} = I_o * e^{(V_g - V_{th}) / kV_t} = w * I_o * e^{(V_g) / kV_t}$$

式中、 $w = e^{(-V_{th}) / kV_t}$ である。

40

【0101】

メモリセル（基準メモリセル又は周辺メモリセルなど）又はトランジスタを使用して入力電流を入力電圧に変換するI-Vログ変換器を使用した場合：

$$V_g = k * V_t * \log [I_{ds} / w p * I_o]$$

式中、 $w p$ は、基準又は周辺メモリセルの w である。

【0102】

ベクトル行列乗算器VMMアレイとして使用されるメモリアレイについて、出力電流は以下である：

$$I_{out} = w a * I_o * e^{(V_g) / kV_t}、すなわち$$

50

$$I_{out} = (w_a / w_p) * I_{in} = W * I_{in}$$

$$W = e^{(V_{thp} - V_{tha}) / kV_t}$$

式中、メモリアレイの各メモリセルの $w_a = w$ である。

【0103】

ワード線又は制御ゲートは、入力電圧のためのメモリセルの入力として使用することができる。

【0104】

代替的に、本明細書に記載されたVMMアレのフラッシュメモリセルは、線形領域において動作するように構成することができる。

$$I_{ds} = \beta * (V_{gs} - V_{th}) * V_{ds}; \beta = u * C_{ox} * W / L$$

$$W = \alpha (V_{gs} - V_{th})$$

10

【0105】

ワード線又は制御ゲート又はビット線又はソース線は、線形領域内において動作するメモリセルの入力として使用することができる。ビット線又はソース線は、メモリセルの出力として使用することができる。

【0106】

I-V線形変換器用に、線形領域において動作するメモリセル（基準メモリセル又は周辺メモリセルなど）又はトランジスタ又は抵抗器を使用して、入出力電流を入出力電圧に線形変換することができる。

【0107】

20

図9のVMMアレ32のための他の実施形態は、参照により本明細書に組み込まれる米国特許出願第15/826,345号に記載されている。上記出願に記載されているように、ソース線又はビット線は、ニューロン出力（電流和出力）として使用することができる。代替的に、本明細書に記載されたVMMアレのフラッシュメモリセルは、飽和領域において動作するように構成することができる。

$$I_{ds} = \alpha_1 / 2 * \beta * (V_{gs} - V_{th})^2; \beta = u * C_{ox} * W / L$$

$$W = \alpha (V_{gs} - V_{th})^2$$

【0108】

ワード線、制御ゲート、又は消去ゲートは、飽和領域内において動作するメモリセルの入力として使用することができる。ビット線又はソース線は、出力ニューロンの出力として使用することができる。

30

【0109】

代替的に、本明細書に記載されるVMMアレのフラッシュメモリセルは、全ての領域又はそれらの組み合わせ（サブスレッショルド、線形、又は飽和）で使用することができる。

【0110】

図12は、図2に示されるメモリセル210に特に適したニューロンVMMアレ1200を示し、入力層と次の層との間のシナプスとして利用される。VMMアレ1200は、不揮発性メモリセルのメモリアレイ1203と、第1の不揮発性基準メモリセルの基準アレ1201と、第2の不揮発性基準メモリセルの基準アレ1202と、を備える。アレの列方向に配置された基準アレ1201及び1202は、端子BLR0、BLR1、BLR2、及びBLR3に流入する電流入力を電圧入力WL0、WL1、WL2、及びWL3に変換するように機能する。実際には、第1及び第2の不揮発性基準メモリセルは、電流入力が流入する状態で、マルチプレクサ1214（一部のみ示す）を通してダイオード接続される。基準セルは、標的基準レベルに調整（例えば、プログラム）される。標的基準レベルは、基準ミニアレマトリックス（図示せず）によって提供される。

40

【0111】

メモリアレイ1203は、2つの目的を果たす。第1に、VMMアレ1200により使用される重みを、それぞれのメモリセルに格納する。第2に、メモリアレイ1203は、メモリセルアレ1203に格納された重みによって、入力（すなわち、端子BLR0

50

、B L R 1、B L R 2、及びB L R 3に提供された電流入力であり、これを基準アレイ1201及び1202が入力電圧に変換して、ワード線W L 0、W L 1、W L 2、及びW L 3に供給する。)を有効に乗算して、次いで、全ての結果(メモリセル電流)を加算して、それぞれのビット線(B L 0～B L N)上の出力を生成し、この出力は次の層への入力又は最後の層への入力となる。メモリアレイ1203が乗算及び加算の関数を実行することで、別個の乗算及び加算の論理回路の必要性はなくなり、また、電力効率も良い。ここで、電圧入力はワード線W L 0、W L 1、W L 2、及びW L 3に提供され、出力は、読み出し(推論)動作中にそれぞれのビット線B L 0～B L Nに現れる。各々のビット線B L 0～B L Nに配置された電流は、その特定のビット線に接続された全ての不揮発性メモリセルからの電流の合計関数を実行する。

10

【0112】

表5は、VMMアレイ1200の動作電圧を示す。表中の列は、選択セルのワード線、非選択セルのワード線、選択セルのビット線、非選択セルのビット線、選択セルのソース線、及び非選択セルのソース線に加えられる電圧を示す。行は、読み出し、消去、及びプログラムの動作を示す。

表5：図12のVMMアレイ1200の動作

【表5】

	WL	WL-非選択	BL	BL-非選択	SL	SL-非選択
読み出し	1～3.5V	-0.5V/0V	0.6～2V (Iニューロン)	0.6V～2V/0V	0V	0V
消去	約5～13V	0V	0V	0V	0V	0V
プログラム	1～2V	-0.5V/0V	0.1～3uA	Vinh約2.5V	4～10V	0～1V/FLT

20

【0113】

図13は、図2に示されるメモリセル210に特に適したニューロンVMMアレイ1300を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。VMMアレイ1300は、不揮発性メモリセルのメモリアレイ1303と、第1の不揮発性基準メモリセルの基準アレイ1301と、第2の不揮発性基準メモリセルの基準アレイ1302と、を備える。基準アレイ1301及び1302は、VMMアレイ1300の行方向に延びる。VMMアレイは、VMMアレイ1300においてワード線が垂直方向に延びることを除いて、VMM1000に類似している。ここで、入力はワード線(W L A 0、W L B 0、W L A 1、W L B 2、W L A 2、W L B 2、W L A 3、W L B 3)に提供され、出力は、読み出し動作中にソース線(S L 0、S L 1)に現れる。各ソース線に加えられる電流は、その特定のソース線に接続されたメモリセルからの全ての電流の合計関数を実行する。

30

【0114】

表6は、VMMアレイ1300の動作電圧を示す。表中の列は、選択セルのワード線、非選択セルのワード線、選択セルのビット線、非選択セルのビット線、選択セルのソース線、及び非選択セルのソース線に加えられる電圧を示す。行は、読み出し、消去、及びプログラムの動作を示す。

40

表6：図13のVMMアレイ1300の動作

50

【表 6】

	WL	WL－非選択	BL	BL－非選択	SL	SL－非選択
読み出し	1～3.5V	－0.5V／0V	0.6～2V	0.6V～2V／0V	約0.3～1V (ニューロン)	0V
消去	約5～13V	0V	0V	0V	0V	SL－inhibit (約4～8V)
プログラム	1～2V	－0.5V／0V	0.1～3μA	Vinh約2.5V	4～10V	0～1V／FLT

【0115】

図14は、図3に示されるメモリセル310に特に適したニューロンVMMアレイ1400を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。VMMアレイ1400は、不揮発性メモリセルのメモリアレイ1403と、第1の不揮発性基準メモリセルの基準アレイ1401と、第2の不揮発性基準メモリセルの基準アレイ1402と、を備える。基準アレイ1401及び1402は、端子BLR0、BLR1、BLR2、及びBLR3に流入する電流入力を電圧入力CG0、CG1、CG2、及びCG3に変換するように機能する。実際には、第1及び第2の不揮発性基準メモリセルは、電流入力がBLR0、BLR1、BLR2、及びBLR3を通して流入する状態で、マルチプレクサ1412（一部のみ示す）を通してダイオード接続される。マルチプレクサ1412は、読み出し動作中に第1及び第2の不揮発性基準メモリセルの各々のビット線（BLR0など）上の一定電圧を確実にするために、それぞれのマルチプレクサ1405及びカスコーディングトランジスタ1404を各々含む。基準セルは、標的基準レベルに調整される。

【0116】

メモリアレイ1403は、2つの目的を果たす。第1に、VMMアレイ1400によって使用される重みを格納する。第2に、メモリアレイ1403は、メモリセルアレイに格納された重みによって、入力（端子BLR0、BLR1、BLR2、及びBLR3に提供された電流入力であり、基準アレイ1401及び1402がこれらの電流入力を入力電圧に変換して、制御ゲート（CG0、CG1、CG2、及びCG3）に供給する）を有効に乗算して、次いで、全ての結果（セル電流）を加算して出力を生成し、この出力はBL0～BLNに出現し、次の層への入力又は最後の層への入力となる。メモリアレイが乗算及び加算の関数を実行することで、別個の乗算及び加算の論理回路の必要性がなくなり、また、電力効率も良い。ここで、入力は制御ゲート線（CG0、CG1、CG2、及びCG3）に提供され、出力は、読み出し動作中にビット線（BL0～BLN）に現れる。各ビット線に加えられる電流は、その特定のビット線に接続されたメモリセルからの全ての電流の合計関数を実行する。

【0117】

VMMアレイ1400は、メモリアレイ1403内の不揮発性メモリセルの一方向調整を実装する。すなわち、各不揮発性メモリセルは消去され、次いで、浮遊ゲート上の所望の電荷に達するまで部分的にプログラムされる。これは、例えば、以下に記載される精密プログラミング技術を使用して実行することができる。過度に多くの電荷が浮遊ゲートに加えられる場合（誤った値がセルに格納されるなど）、セルは消去されなければならない。示されるように、同じ消去ゲート（EG0又はEG1など）を共有する2つの行は、一緒に消去される必要があり（ページ消去として知られる）、その後、各セルは、浮遊ゲート上の所望の電荷に達するまで部分的にプログラムされる。

【0118】

表7は、VMMアレイ1400の動作電圧を示す。表中の列は、選択セルのワード線、非選択セルのワード線、選択セルのビット線、非選択セルのビット線、選択セルの制御ゲ

ート、選択セルと同じセクタ内の非選択セルの制御ゲート、選択セルとは異なるセクタ内の非選択セルの制御ゲート、選択セルの消去ゲート、非選択セルの消去ゲート、選択セルのソース線、及び非選択セルのソース線に加えられる電圧を示す。行は、読み出し、消去、及びプログラムの動作を示す。

表 7：図 14 の VMM アレイ 1400 の動作

【表 7】

	WL	WL-非選択	BL	BL-非選択	CG	CG-非選択、 同じセクタ	CG- 非選択	EG	EG- 非選択	SL	SL- 非選択
読み出し	1.0~2V	-0.5V/0V	0.6~2V (1ニューロン)	0V	0~2.6V	0~2.6V	0~2.6V	0~2.6V	0~2.6V	0V	0V
消去	0V	0V	0V	0V	0V	0~2.6V	0~2.6V	5~12V	0~2.6V	0V	0V
プログラム	0.7~1V	-0.5V/0V	0.1~1uA	Vinh(1~2V)	4~11V	0~2.6V	0~2.6V	4.5~5V	0~2.6V	4.5~5V	0~1V

10

【0119】

図 15 は、図 3 に示されるメモリセル 310 に特に適したニューロン VMM アレイ 1500 を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。VMM アレイ 1500 は、不揮発性メモリセルのメモリアレイ 1503 と、基準アレイ 1501 又は第 1 の不揮発性基準メモリセルと、第 2 の不揮発性基準メモリセルの基準アレイ 1502 と、を備える。EG 線 EGR0、EG0、EG1、及び EGR1 は垂直に延び、CG 線 CG0、CG1、CG2、及び CG3 並びに SL 線 WL0、WL1、WL2、及び WL3 は水平に延びる。VMM アレイ 1500 は、VMM アレイ 1400 が双方向調整を実装することを除いて VMM アレイ 1400 に類似しており、各個々のセルは、個別の EG 線の使用により、浮遊ゲート上の所望の電荷量に達するために、完全に消去され、部分的にプログラムされ、必要に応じて部分的に消去することができる。示されるように、基準アレイ 1501 及び 1502 は、端子 BLR0、BLR1、BLR2 及び BLR3 内の入力電流を制御ゲート電圧 CG0、CG1、CG2 及び CG3 に変換し（マルチプレクサ 1514 を介したダイオード接続された基準セルの作用を通じて）、これらの電圧は行方向でメモリセルに印加される。電流出力（ニューロン）は、ビット線 BL0~BLN 中にあり、各ビット線は、その特定のビット線に接続された不揮発性メモリセルからの全ての電流を合計する。

20

【0120】

表 8 は、VMM アレイ 1500 の動作電圧を示す。表中の列は、選択セルのワード線、非選択セルのワード線、選択セルのビット線、非選択セルのビット線、選択セルの制御ゲート、選択セルと同じセクタ内の非選択セルの制御ゲート、選択セルとは異なるセクタ内の非選択セルの制御ゲート、選択セルの消去ゲート、非選択セルの消去ゲート、選択セルのソース線、及び非選択セルのソース線に加えられる電圧を示す。行は、読み出し、消去、及びプログラムの動作を示す。

表 8：図 15 の VMM アレイ 1500 の動作

【表 8】

	WL	WL-非選択	BL	BL-非選択	CG	CG-非選択、 同じセクタ	CG- 非選択	EG	EG- 非選択	SL	SL- 非選択
読み出し	1.0~2V	-0.5V/0V	0.6~2V (1ニューロン)	0V	0~2.6V	0~2.6V	0~2.6V	0~2.6V	0~2.6V	0V	0V
消去	0V	0V	0V	0V	0V	4~9V	0~2.6V	5~12V	0~2.6V	0V	0V
プログラム	0.7~1V	-0.5V/0V	0.1~1uA	Vinh(1~2V)	4~11V	0~2.6V	0~2.6V	4.5~5V	0~2.6V	4.5~5V	0~1V

40

【0121】

図 24 は、図 2 に示されるメモリセル 210 に特に適したニューロン VMM アレイ 2400 を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。VMM アレイ 2400 では、入力 INPUT₀、...、INPUT_N は、ビット線 BL₀、...、BL_N でそれぞれ受信され、出力 OUTPUT₁、OUTPUT₂、OUTPUT₃、及び OUTPUT₄ は、ソース線 SL₀、SL₁、SL₂、及び SL₃ でそれぞれ生成され

50

る。

【0122】

図25は、図2に示されるメモリセル210に特に適したニューロンVMMアレイ2500を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。この例では、入力 $INPUT_0$ 、 $INPUT_1$ 、 $INPUT_2$ 、及び $INPUT_3$ は、ソース線 SL_0 、 SL_1 、 SL_2 、及び SL_3 でそれぞれ受信され、出力 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ は、ビット線 BL_0 、 $\dots$ 、 BL_N で生成される。

【0123】

図26は、図2に示されるメモリセル210に特に適したニューロンVMMアレイ2600を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。この例では、入力 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ は、ワード線 WL_0 、 $\dots$ 、 WL_M でそれぞれ受信され、出力 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ は、ビット線 BL_0 、 $\dots$ 、 BL_N で生成される。

10

【0124】

図27は、図3に示されるメモリセル310に特に適したニューロンVMMアレイ2700を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。この例では、入力 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ は、ワード線 WL_0 、 $\dots$ 、 WL_M でそれぞれ受信され、出力 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ は、ビット線 BL_0 、 $\dots$ 、 BL_N で生成される。

【0125】

図28は、図4に示されるメモリセル410に特に適したニューロンVMMアレイ2800を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。この例では、入力 $INPUT_0$ 、 $\dots$ 、 $INPUT_n$ は、制御ゲート線 CG_0 、 $\dots$ 、 CG_N でそれぞれ受信され、出力 $OUTPUT_1$ 及び $OUTPUT_2$ は、ソース線 SL_0 及び SL_1 で生成される。

20

【0126】

図29は、図4に示されるメモリセル410に特に適したニューロンVMMアレイ2900を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。この例では、入力 $INPUT_0$ 、 $\dots$ 、 $INPUT_N$ は、ビット線 BL_0 、 $\dots$ 、 BL_N にそれぞれ結合されたビット線制御ゲート2901-1、2901-2、 $\dots$ 、2901-(N-1)、及び2901-Nのゲートでそれぞれ受信される。例示的な出力 $OUTPUT_1$ 及び $OUTPUT_2$ が、ソース線 SL_0 及び SL_1 で生成される。

30

【0127】

図30は、図3に示されるメモリセル310、図5に示されるメモリセル510、及び図7に示されるメモリセル710に特に適したニューロンVMMアレイ3000を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。この例では、入力 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ は、ワード線 WL_0 、 $\dots$ 、 WL_M で受信され、出力 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ は、ビット線 BL_0 、 $\dots$ 、 BL_N でそれぞれ生成される。

【0128】

図31は、図3に示されるメモリセル310、図5に示されるメモリセル510、及び図7に示されるメモリセル710に特に適したニューロンVMMアレイ3100を示し、入力層と次の層との間でシナプス及びニューロンの一部として利用される。この例では、入力 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ は、制御ゲート線 CG_0 、 $\dots$ 、 CG_M で受信される。出力 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ は、ソース線 SL_0 、 $\dots$ 、 SL_N でそれぞれ生成され、各ソース線 SL_i は、列*i*内の全てのメモリセルのソース線端子に結合される。

40

【0129】

図32は、VMMシステム3200を示す。VMMシステム3200は、VMMアレイ3201（VMM900、1000、1100、1200、及び1320などの先で考察

50

されるVMM設計、又は他のVMM設計のいずれかに基づくことができる）、低電圧行デコーダ3202、高電圧行デコーダ3203、基準セル低電圧列デコーダ3204（列方向に示されており、これは行方向の変換を出力する入力を提供することを意味する）、ビット線マルチプレクサ3205、制御論理3206、アナログ回路3207、ニューロン出力ブロック3208、入力VMM回路ブロック3209、プリデコーダ3210、試験回路3211、消去プログラム制御論理E P C T L 3212、アナログ及び高電圧生成回路3213、ビット線P Eドライバ3214、冗長アレイ3215及び3216、N V Rセクタ3217、並びに基準セクタ3218を備える。入力回路ブロック3209は、外部入力からメモリアレイの入力端子へのインターフェースとして機能する。ニューロン出力ブロック3208は、メモリアレイ出力から外部インターフェースへのインターフェースとして機能する。

10

【0130】

低電圧行デコーダ3202は、読み出し動作及びプログラム動作のためのバイアス電圧を提供し、高電圧行デコーダ3203にデコード信号を提供する。高電圧行デコーダ3203は、プログラム動作及び消去動作のための高電圧バイアス信号を提供する。任意選択的な基準セル低電圧列デコーダ3204は、基準セルのデコード関数を提供する。ビット線P Eドライバ3214は、プログラム動作、検証動作、及び消去動作中のビット線の制御関数を提供する。アナログ及び高電圧生成回路3213は、様々なプログラム動作、消去動作、プログラム検証動作、及び読み出し動作に必要な複数の電圧を提供する共有バイアスブロックである。任意選択的な冗長アレイ3215及び3216は、欠陥のあるアレイ部分を置き換えるためのアレイ冗長性を提供する。任意選択的なN V R（不揮発性レジスタ、情報セクタとも知られている）セクタ3217は、限定されることなく、ユーザ情報、デバイスID、パスワード、セキュリティキー、トリムビット、構成ビット、製造情報を格納するために使用されるアレイセクタであるセクタである。

20

【0131】

図33は、アナログニューロメモリシステム3300を示す。アナログニューロメモリシステム3300は、マクロブロック3301a、3301b、3301c、3301d、3301e、3301f、3301g及び3301hと、ニューロン出力（加算器回路及びサンプルホールドS/H回路など）ブロック3302a、3302b、3302c、3302d、3302e、3302f、3302g、及び3302hと、入力回路ブロック3303a、3303b、3303c、3303d、3303e、3303f、3303g、及び3304hと、を備える。マクロブロック3301a、3301b、3301c、3301d、3301e、及び3301fの各々は、VMMアレイを含むVMMサブシステムである。ニューロメモリサブシステム3333は、マクロブロック3301、入力ブロック3303、及びニューロン出力ブロック3302を含む。ニューロメモリサブシステム3333は、それ自体のデジタル制御ブロックを有してもよい。

30

【0132】

アナログニューロメモリシステム3300は、システム制御ブロック3304、アナログ低電圧ブロック3305、及び高電圧ブロック3306を更に備える。

システム制御ブロック3304は、一般的な制御機能動作及び演算動作を処理するためのARM/MIPS/RISC_Vコアなどのマイクロコントローラコアを含んでもよい。システム制御ブロック3304はまた、単一の命令を用いて複数のデータ上で動作するSIMD（単一命令複数データ）ユニットを含んでもよい。DSPコアを含んでもよい。限定されることなく、プーリング、平均化、min、max、softmax、add、subtract、multiply、divide、対数、逆対数、ReLU、シグモイド、tanh、データ圧縮などの関数を実行するためのハードウェア又はソフトウェアを含んでもよい。活性化近似器／量子化器／正規化器などの関数を実行するハードウェア又はソフトウェアを含んでもよい。入力データ近似器／量子化器／正規化器などの関数を実行する能力を含んでもよい。活性化近似器／量子化器／正規化器の関数を実行するハードウェア又はソフトウェアを含んでもよい。ニューロメモリサブシステム3333の制御

40

50

ブロックは、マイクロコントローラコア、SIMDコア、DSPコア、及び他の関数ユニットなどのシステム制御ブロック3304の類似の要素を含んでもよい。

【0133】

一実施形態では、ニューロン出力ブロック3302a、3302b、3302c、3302d、3302e、3302f、3302g、及び3302hは、構成可能な長い相互接続を駆動することができる低インピーダンス出力型バッファ（例えば、オペアンプ）回路を各々含む。一実施形態では、入力回路ブロック3303a、3303b、3303c、3303d、3303e、3303f、3303g、及び3303hは、加算の高インピーダンス電流出力を各々提供する。別の実施形態では、ニューロン出力ブロック3302a、3302b、3302c、3302d、3302e、3302f、3302g、及び3302hは、活性化回路を各々含み、この場合、出力を駆動するために追加の低インピーダンスバッファが必要とされる。

10

【0134】

別の実施形態では、ニューロン出力ブロック3302a、3302b、3302c、3302d、3302e、3302f、3302g、及び3302hは、アナログ信号の代わりにデジタルビットを出力するアナログデジタル変換ブロックを各々含む。この実施形態では、入力回路ブロック3303a、3303b、3303c、3303d、3303e、3303f、3303g、及び3303hは、それぞれのニューロン出力ブロックからデジタルビットを受信してデジタルビットをアナログ信号に変換するデジタルアナログ変換ブロックを各々含む。

20

【0135】

したがって、ニューロン出力ブロック3302a、3302b、3302c、3302d、3302e、3302f、3302g、及び3302hは、マクロブロック3301a、3301b、3301c、3301d、3301e、及び3301fからの出力電流を受信し、任意選択的に、その出力電流をアナログ電圧、デジタルビット、又は、各パルスの幅若しくはパルス数が出力電流の値に应答して変動する、1つ以上のデジタルパルスに変換する。同様に、入力回路ブロック3303a、3303b、3303c、3303d、3303e、3303f、3303g、及び3303hは、任意選択的に、アナログ電流、アナログ電圧、デジタルビット、又は、各パルスの幅若しくはパルス数が出力電流の値に应答して変動する、デジタルパルスを受信し、マクロブロック3301a、3301b、3301c、3301d、3301e、及び3301fにアナログ電流を提供する。入力回路ブロック3303a、3303b、3303c、3303d、3303e、3303f、3303g、及び3303hは、任意選択的に、電圧電流変換器、入力信号内のデジタルパルス数若しくは入力信号内のデジタルパルスの幅を計数するためのアナログ若しくはデジタルカウンタ、又はデジタルアナログ変換器を含む。

30

長・短期記憶

【0136】

先行技術は、長・短期メモリ（long short-term memory、LSTM）として知られる概念を含む。LSTMユニットは、しばしば、ニューラルネットワーク内で使用される。LSTMによって、ニューラルネットワークは所定の任意の期間にわたって情報を記憶し、後続の動作においてその情報を使用することができる。従来のLSTMユニットは、セル、入力ゲート、出力ゲート、及び忘却ゲートを含む。3つのゲートは、セル内及びセル外への情報の流れ、及び情報がLSTM内で記憶される期間を調整する。VMMは、LSTMユニットにおいて特に有用である。

40

【0137】

図16は、例示的なLSTM1600を示す。この例におけるLSTM1600は、セル1601、1602、1603及び1604を含む。セル1601は、入力ベクトル x_0 を受け取り、出力ベクトル h_0 及びセル状態ベクトル c_0 を生成する。セル1602は、入力ベクトル x_1 と、セル1601からの出力ベクトル（隠れ状態） h_0 と、セル1601からのセル状態 c_0 と、を受け取り、出力ベクトル h_1 と、セル状態ベクトル c_1 と、を生成

50

する。セル1603は、入力ベクトル x_2 と、セル1602からの出力ベクトル（隠れ状態） h_1 と、セル1602からのセル状態 c_1 と、を受け取り、出力ベクトル h_2 と、セル状態ベクトル c_2 と、を生成する。セル1604は、入力ベクトル x_3 と、セル1603からの出力ベクトル（隠れ状態） h_2 と、セル1603からのセル状態 c_2 と、を受け取り、出力ベクトル h_3 を生成する。追加のセルも使用可能であり、4つのセルを有するLSTMは、単なる例である。

【0138】

図17は、図16のセル1601、1602、1603及び1604に使用可能なLSTMセル1700の例示的な実装形態を示す。LSTMセル1700は、入力ベクトル $x(t)$ と、先行するセルからのセル状態ベクトル $c(t-1)$ と、先行するセルからの出力ベクトル $h(t-1)$ と、を受け取り、セル状態ベクトル $c(t)$ 及び出力ベクトル $h(t)$ を生成する。

【0139】

LSTMセル1700は、シグモイド関数デバイス1701、1702、及び1703を備え、各々が0～1の数を適用することで、入力ベクトルの各成分が出力ベクトルに寄与する程度を制御する。LSTMセル1700はまた、入力ベクトルに双曲線正接関数を適用するための $\tanh$ デバイス1704及び1705と、2つのベクトルを乗算するための乗算器デバイス1706、1707、及び1708と、2つのベクトルを加算するための加算器デバイス1709と、を含む。出力ベクトル $h(t)$ は、システム内の次のLSTMセルに提供することができ、又は他の目的でアクセスすることができる。

【0140】

図18は、LSTMセル1700の実装形態の例であるLSTMセル1800を示す。読者の便宜のために、LSTMセル1700からの同じ採番方法が、LSTMセル1800で使用される。シグモイド関数デバイス1701、1702及び1703、並びに $\tanh$ デバイス1704は各々、複数のVMMアレイ1801及び活性化回路ブロック1802を含む。したがって、VMMアレイは、特定のニューラルネットワークシステムで使用されるLSTMセルにおいて特に有用であることが理解できる。乗算器デバイス1706、1707、及び1708、並びに加算器デバイス1709は、デジタル様態において、又はアナログ様態において実装される。活性化関数ブロック1802は、デジタル様態において、又はアナログ様態において実装され得る。

【0141】

LSTMセル1800の代替例（及びLSTMセル1700の実装形態の別の例）を図19に示す。図19では、シグモイド関数デバイス1701、1702及び1703、並びに $\tanh$ デバイス1704は、同じ物理ハードウェア（VMMアレイ1901及び活性化関数ブロック1902）を、時分割多重化された方式で共有する。LSTMセル1900はまた、2つのベクトルを乗算するための乗算器デバイス1903と、2つのベクトルを加算するための加算器デバイス1908と、（活性化回路ブロック1902を含む） $\tanh$ デバイス1705と、シグモイド関数ブロック1902から出力される値 $i(t)$ を格納するレジスタ1907と、マルチプレクサ1910を介して乗算器デバイス1903から出力される値 $f(t) * c(t-1)$ を格納するレジスタ1904と、マルチプレクサ1910を介して乗算器デバイス1903から出力される値 $i(t) * u(t)$ を格納するレジスタ1905と、マルチプレクサ1910を介して乗算器デバイス1903から出力される値 $o(t) * c \sim (t)$ を格納する、レジスタ1906と、マルチプレクサ1900とを含む。

【0142】

LSTMセル1800が複数のVMMアレイ1801とそれぞれの活性化関数ブロック1802のセットを複数含むのに対し、LSTMセル1900は、LSTMセル1900の実施形態において複数の層を表すために使用されるVMMアレイ1901と活性化関数ブロック1902のセットを1つのみ含む。LSTMセル1900は、LSTMセル1800と比較して、VMM及び活性化関数ブロックのために必要とするスペースは1/4で

済むので、LSTMセル1900は、LSTM1800よりも必要とするスペースが少ない。

【0143】

LSTMユニットは典型的には複数のVMMアレイを備え、これらは各々、加算器及び活性化回路ブロック及び高電圧生成ブロックなどの、VMMアレイの外側の特定の回路ブロックによって提供される機能を必要とすることが更に理解され得る。各VMMアレイのための別個の回路ブロックを提供することは、半導体デバイス内にかなりの量のスペースを必要とし、幾分非効率的であろう。

ゲート付き回帰型ユニット

【0144】

アナログVMM実装形態は、ゲート付き回帰型ユニット (gated recurrent unit、GRU) システムに利用することができる。GRUは、回帰型ニューラルネットワーク内のゲート機構である。GRUは、GRUセルが一般にLSTMセルよりも少ない構成要素を含むことを除いて、LSTMに類似している。

【0145】

図20は、例示的なGRU2000を示す。この例におけるGRU2000は、セル2001、2002、2003、及び2004を含む。セル2001は、入力ベクトル x_0 を受け取り、出力ベクトル h_0 を生成する。セル2002は、入力ベクトル x_1 と、セル2001からの出力ベクトル h_0 と、を受け取り、出力ベクトル h_1 を生成する。セル2003は、入力ベクトル x_2 と、セル2002からの出力ベクトル (隠れ状態) h_1 と、を受け取り、出力ベクトル h_2 を生成する。セル2004は、入力ベクトル x_3 と、セル2003からの出力ベクトル (隠れ状態) h_2 と、を受け取り、出力ベクトル h_3 を生成する。追加のセルも使用可能であり、4つのセルを有するGRUは、単なる例である。

【0146】

図21は、図20のセル2001、2002、2003、及び2004に使用可能なGRUセル2100の例示的な実装形態を示す。GRUセル2100は、入力ベクトル $x(t)$ と、先行するGRUセルからの出力ベクトル $h(t-1)$ と、を受け取り、出力ベクトル $h(t)$ を生成する。GRUセル2100は、シグモイド関数デバイス2101及び2102を備え、各々が、出力ベクトル $h(t-1)$ 及び入力ベクトル $x(t)$ からの成分に0~1の数を適用する。GRUセル2100はまた、入力ベクトルに双曲線正接関数を適用するための $\tanh$ デバイス2103と、2つのベクトルを乗算するための乗算器デバイス2104、2105及び2106と、2つのベクトルを加算するための加算器デバイス2107と、1から入力を減算して出力を生成するコンプリメンタリデバイス2108と、を含む。

【0147】

図22は、GRUセル2100の実装形態の例であるGRUセル2200を示す。読者の便宜のために、GRUセル2100からの同じ採番方法が、GRUセル2200で使用される。図22から分かるように、シグモイド関数デバイス2101及び2102、並びに $\tanh$ デバイス2103は各々、複数のVMMアレイ2201及び活性化関数ブロック2202を含む。したがって、VMMアレイは、特定のニューラルネットワークシステムで使用されるGRUセルにおいて特に使用されることが分かる。乗算器デバイス2104、2105、2106、加算器デバイス2107、及びコンプリメンタリデバイス2108は、デジタル様態において、又はアナログ様態において実装される。活性化関数ブロック2202は、デジタル様態において、又はアナログ様態において実装され得る。

【0148】

GRUセル2200の代替例 (及びGRUセル2300の実装形態の別の例) を図23に示す。図23において、GRUセル2300は、VMMアレイ2301及び活性化関数ブロック2302を利用しており、シグモイド関数として構成された場合には、0~1の数を適用することで、入力ベクトルの各成分が出力ベクトルに寄与する程度を制御する。図23では、シグモイド関数デバイス2101及び2102、並びに $\tanh$ デバイス2

10

20

30

40

50

103は、同じ物理ハードウェア（VMMアレイ2301及び活性化関数ブロック2302）を、時分割多重化された方式で共有する。GRUセル2300はまた、2つのベクトルを乗算するための乗算器デバイス2303と、2つのベクトルを加算するための加算器デバイス2305と、1から入力を減算して出力を生成するためのコンプリメンタリデバイス2309と、マルチプレクサ2304と、マルチプレクサ2304を介して乗算器デバイス2303から出力される値 $h(t-1) * r(t)$ を保持するレジスタ2306と、マルチプレクサ2304を介して乗算器デバイス2303から出力される値 $h(t-1) * z(t)$ を保持するレジスタ2307と、マルチプレクサ2304を介して乗算器デバイス2303から出力される値 $h^{\wedge}(t) * (1 - z(t))$ を保持するレジスタ2308と、を含む。

10

【0149】

GRUセル2200が複数のVMMアレイ2201と活性化関数ブロック2202のセットを複数含むのに対し、GRUセル2300は、GRUセル2300の実施形態において複数の層を表すために使用されるVMMアレイ2301と活性化関数ブロック2302のセット1つのみを含む。GRUセル2300は、GRUセル2200と比較して、VMM及び活性化関数ブロックのために必要とするスペースは1/3で済むので、GRUセル2300は、GRUセル2200よりも必要とするスペースが少ない。

【0150】

GRUシステムは典型的には複数のVMMアレイを備え、これらは各々、加算器及び活性化回路ブロック及び高電圧生成ブロックなどの、VMMアレイの外側の特定の回路ブロックによって提供される機能を必要とすることが更に理解され得る。各VMMアレイのための別個の回路ブロックを提供することは、半導体デバイス内にかなりの量のスペースを必要とし、幾分非効率的であろう。

20

【0151】

VMMアレイへの入力は、アナログレベル、バイナリレベル、又はデジタルビット（この場合、デジタルビットを適切な入力アナログレベルに変換するためにDACが必要とされる）であり、出力はアナログレベル、バイナリレベル、又はデジタルビット（この場合、出力アナログレベルをデジタルビットに変換するために出力ADCが必要とされる）であり得る。

【0152】

VMMアレイ内の各メモリセルに関して、各重み w は、単一のメモリセルによって、又は差分セルによって、又は2つのブレンドメモリセル（2つのセルの平均）によって実装することができる。差分セルの場合では、重み w を差分重み（ $w = w_+ - w_-$ ）として実装するために、2つのメモリセルが必要とされる。2つのブレンドメモリセルの場合は、2つのセルの平均として重み w を実装するために2つのメモリセルが必要とされる。

30

高電圧生成回路及び他の回路

【0153】

図34は、VMMシステム3400のブロック図を示す。VMMシステム3400は、VMMアレイ3408、行デコーダ3407、高電圧デコーダ3409、列デコーダ3410、及びビット線ドライバ3411を備える。VMMシステム3400は、チャージポンプ3401、チャージポンプ調整器3402、及び高電圧レベル生成器3403を備える高電圧生成ブロック3412を更に備える。VMMシステム3400は、アルゴリズムコントローラ3404、アナログ回路3405、及び制御論理3406を更に備える。

40

【0154】

図35は、チャージポンプ3401及びチャージポンプ調整器3402に関する更なる詳細を提供する。チャージポンプ3401は、イネーブル信号3501によって制御される。イネーブル信号3501がアサートされないとき、チャージポンプ3401は、その出力の電圧を増加させ続ける。イネーブル信号3501がアサートされるとき、チャージポンプ3401は、その出力の電圧レベルを維持する。チャージポンプ調整器3402は、ダイオード3504、3506、及び3508、並びにダイオード3504、3506

50

、及び3508のそれぞれの1つのカソードに各々結合された抵抗器3505、3507、及び3509の直列接続を含む分圧器構造を備える。その構造内の分圧電圧ノードは、電圧基準を含む別の入力を受信するコンパレータ3503に入力される。チャージポンプ3401からの電圧出力が、電流が流れるようにダイオード3504、3506及び3508を活性化させるのに十分であり、分圧電圧ノードからの電圧が電圧基準を超えると、イネーブル信号がアサートされる。したがって、チャージポンプ調整器3402は、ダイオード3504、3506、及び3508並びに抵抗器3505、3507、及び3509の特性に基づく所望の電圧レベルが達成されるまでチャージポンプ3401を制御する。3つのダイオード及び3つの抵抗器が、分圧器構造の例として示されており、典型的には3つを超えるものが必要である。代替的に、ダイオード及び抵抗器の代わりにコンデンサを実装して、コンパレータ3503への入力を提供するために所望の電圧比を生成することができる。代替的に、適切な比率のコンデンサをダイオード及び抵抗器と並列に接続して、分圧器構造の応答を加速させることができる、

【0155】

図36は、VMMシステム3400の一実施形態であるVMMシステム3600を示す。VMMシステム3600は、高電圧バッファ3601及び調整可能な電流シンク3602を備える。高電圧生成ブロック3412は、高電圧バッファ3601に提供される電圧を生成し、これは次に、その電圧を高電圧デコーダ3409及び調整可能な電流シンク（プログラム補償電流Icomp）3602に提供する。高電圧バッファ3601から調整可能な電流シンク3602によって引き込まれる電流Icompは、例えば、プログラムされるメモリセルの数を補償するために、高電圧バッファ3601内に補償された電圧ドロップ（例えば、プログラムされる1/2/. . . /32IOに対してdVout1/2/. . . /32ドロップ）を誘発するように、かつ高電圧バッファ3601の温度を低下させるように調整することができる。例えば、 $I_{comp} = (\text{プログラムされる\#メモリセル}) * I_{prog} * M$ （式中、Iprog=セルプログラミング電流、及びM=プログラミング動作中のメモリセルのホットキャリア効果による乗数である）。補償Icompは、変動する出力負荷にわたって一定の高電圧出力を維持するために適用される。

【0156】

図37は、高電圧バッファ3701及び調整可能な電流シンク3702と共に使用されるVMMシステム3700の一実施形態を示す。高電圧生成器3412は、高電圧バッファ3701に提供される電圧を生成し、これは次に、その電圧を高電圧デコーダ3409に提供する。高電圧デコーダ3409から調整可能な電流シンク（補償電流）Icomp3702によって引き込まれる電流は、例えば、高電圧デコーダ3409内の電流ドロップ（プログラムされるメモリセルの数の関数として）を低減して、及び／又は、高電圧デコーダ3409の温度を低下させるように調整することができる。例えば、 $I_{comp} = (\text{プログラムされる\#メモリセル}) * I_{prog} * M$ 。Iprog=セルプログラミング電流、M=プログラミング動作中のメモリセルのホットキャリア効果による乗数である。補償Icompは、変動する出力負荷にわたって一定の高電圧出力を維持するために適用される。

【0157】

図38は、ここでは演算増幅器である高電圧バッファ3801と共に使用されるVMMシステム3800を示す。高電圧生成器3412は、高電圧バッファ3801に提供される電圧を生成し、これは次に、その電圧を高電圧デコーダ3409に提供する。高電圧デコーダ3409からの出力（例えば、出力はデコーダ内のHV電圧のフィードバックインジケータである）は、高電圧バッファ3801への入力として提供され、次いで、高電圧バッファ3801は、閉ループ演算増幅器として動作する。閉ループ補償は、変動する出力負荷にわたって一定の高電圧出力を維持するために適用される。

【0158】

図39は、例えば、各VMMシステム内のVMMアレイの追加として、VMMシステム2400、2600、2700、又は2800と共に使用されるプログラム電流補償ブロ

10

20

30

40

50

ック3900を示す。ここで、ダミープログラムビット線（プログラム可能ダミーアレイ）には、32ビット線の各グループが提供されている。例えば、グループ3901はダミービット線3903を含み、グループ3902はダミービット線3904を含む。これらのダミービット線3903及び3904は、それぞれ、グループ3901及び3902内の1つ以上の他のビットがプログラムされていない場合に（ビット線プログラミング電流を提供するために）オンにすることができる。これにより、プログラミング動作中に引き込まれる電流が、ダミービット線3903及び3904を使用しない場合より一定に維持される。プログラムダミーアレイ補償スキームは、変動する出力負荷にわたって一定の高電圧出力を維持するために適用される。

【0159】

図40は、高電圧デコーダ3409を実装するために使用することができる高電圧デコーダブロック4000の例を示す。ここで、ソース線4005は、VMMアレイ3408内の1つ又は2つの行に結合されている。NMOSトランジスタ4001、4002、4003、及び4004は、図示のようにソース線4005に結合されている。HV電源4010は、HVバッファ3601、3701又は3801などのHVバッファから提供され、HV比較信号4011は、図38に示されるようなものである。

【0160】

図41は、試験回路4100を示す。試験回路4100は、イネーブル信号ENを受信する高電圧トランスミッタ4101を備える。高電圧トランスミッタは、NMOSトランジスタ4102及びNMOSカスコードトランジスタ4103に高電圧イネーブル信号を提供する。NMOSトランジスタ4102の一方の端子は外部試験パッドに接続し、NMOSトランジスタ4103の一方の端子は、VMMシステム3400内の内部ノードに結合される。この回路は、例えば、電圧較正プロセス中に使用することができる。

【0161】

図42は、ここでは高電圧生成回路4200、制御論理ブロック4201、アナログ回路ブロック4202、及び試験ブロック4203を備える高電圧生成ブロック3412の一実施形態を示す。高電圧生成回路4200は、チャージポンプ及び調整器4204と、高電圧インクリメンタ4205と、高電圧演算増幅器4206と、を備える。高電圧インクリメンタ4205の出力の電圧は、以下で更に説明するように、高電圧インクリメンタ4205内のトランジスタに送られるトリム信号に基づいて制御され得る。制御論理ブロック4201は、制御論理入力を受信し、制御論理出力を生成する。アナログ回路ブロック4202は、基準電圧Vrefを受け取り、他の場所で使用されるようにバイアス信号iBiasを付与するために使用可能な電流を生成するための電流バイアス発生器4207を含む。アナログ回路ブロック4202はまた、トリムビットのセットTRBIT_WLを受け取り、各種動作中にワード線に印加する電圧を生成するための電圧発生器4208を含む。試験ブロック4203は、試験パッドMONHV_PADで信号を受信し、試験中に監視するための様々な信号を出力する。

【0162】

図43は、高電圧生成ブロック3412の別の実施形態を示す。ここでは、高電圧生成ブロックは、チャージポンプ及び調整器4301と、高電圧（high voltage、HV）インクリメンタ4303と、高電圧演算増幅器4302と、を備える。高電圧インクリメンタ4303の出力の電圧は、高電圧インクリメンタ4303内のトランジスタのゲートに送られる信号に基づいて制御することができる。HVインクリメンタ4303は、接地からチャージポンプ4301の出力に直列に接続された抵抗器ストリング4315iを備える。スイッチ4310a、4310b、4310zのネットワークは、インクリメント方式でストリングに沿って電圧レベルを脱多重化する（mux out）ために使用される。トランジスタのゲートは、高電圧レベルシフタ（high voltage level shifter、HVLS）4320によってイネーブル／ディスエーブルされ、これは次に、デジタル制御入力によってイネーブル／ディスエーブルされる。HVLS4320は、デジタル電圧レベル、例えば、1Vを高電圧レベル、例えば、12Vに変換するために使用される。例えば、抵抗ス

10

20

30

40

50

トリング 4315 i は、10 mV の電圧インクリメント（1 つの抵抗器にわたる電圧）で 3 V ~ 10 V の電圧レベルを提供する。したがって、抵抗ストリングの出力 VHVROUT は、10 mV のインクリメントで 3 V ~ 10 V の値を有する。HV 演算増幅器 4302 は、この VHVROUT インクリメント電圧をバッファするために使用される。最高電圧、例えば、10 V で必要とされる HV 電圧により、この値における又は近い PMOS スイッチは、耐圧（breakdown voltage、BV）に近い状態に関連する漏れを見ることになる。これは、小さいインクリメント電圧、例えば、10 mV の精度に影響を及ぼす。したがって、この BV 漏れを克服するために改善が必要とされる。まず、PMOS スイッチの位置に従って、HVL S 回路の供給をタップする。例えば、4 ~ 6 V のストリング位置での PMOS スイッチの場合、HVL S の高供給は、6 V（より典型的な 12 V 供給の代わりに）である。更に、低供給は、（より典型的な接地値の代わりに）4 V であり得る。これにより、この抵抗器ストリング位置において、PMOS スイッチ全体にわたる電圧ストレスが低減される。VHVROUT に接続された PMOS スイッチの場合、適応 HV バイアスと直列の 2 つが、BV ストレスを回避するために、それぞれの mux 経路をディスエーブルするために必要とされ、例えば、PMOS 4310 e / 4310 f 及び 4310 y / 4310 z が示されている。PMOS 4310 f 及び 4310 z は、mux 経路ディスエーブル化のために使用される。例えば、PMOS 4310 f 及び 4310 z のゲートは、10 V の mux 経路をディスエーブルするために、高電圧（例えば、10 V）である。オフ状態では、PMOS 4310 e のゲートは、好ましくは、BV 漏れを低減するためのカスコーディングのために約 6 V であり、PMOS 4310 f のゲートは 10 V である。オン状態では、PMOS 4310 f のゲートは約 6 V であり得、PMOS 4310 f のゲートは、それぞれのストリングから VHVROUT への 8 V ~ 10 V（例として）の mux 経路を通過させるために < 6 V であり得る。オフ状態では、PMOS 4310 y のゲートは、BV 漏れを低減するためのカスコーディングのために約 6 V であり、PMOS 4310 z のゲートは 10 V である。オン状態では、PMOS 4310 y のゲートは約 0 V であり得、PMOS 4310 z のゲートは、ストリングから VHVROUT への 3 V ~ 5 V（例として）の mux 経路を通過させるために < 0 V であり得る。

【0163】

図 44 は、高電圧生成ブロック 3412 の別の実施形態を示す。高電圧生成ブロック 3412 は、高電圧演算増幅器 4403、SC（スイッチキャップ）ネットワーク 4402、及び SC ネットワーク 4401 を備える。SC ネットワーク 4402 は、調整可能なコンデンサ 4404 を備える。SC ネットワーク 4401 は、スイッチ 4405、4407、4408、及び 4409、並びに調整可能なコンデンサ 4406 を備える。HV レベル（例えば、10 ~ 13 V）で動作する高電圧レベルシフタ（HVL S）回路が、SC ネットワーク 4402 用のコンデンサ 4404 を調整するために必要とされる。SC ネットワーク 4401 には、IO 電圧（例えば、1.8 V、2.5 V）又はコア電圧（例えば、1.2 V）のスイッチング回路が必要とされる。

【0164】

図 45 は、図 44 の高電圧演算増幅器 4403 に使用することができる高電圧演算増幅器 4500 を示す。高電圧演算増幅器 4500 は、示された配置で示される構成要素を備える。HV カスコードバイアスノード VCASP、VCASN1、及び VCASN2 は、トランジスタにわたる最大ストレス電圧ドロップを最小化するために、電圧値が出力電圧 VOUT に依存するように適応的な方式で実装される。例えば、ノード電圧 VOUT が高いとき、VCASN2 は高く、VCASN1 は低い。

【0165】

図 46 は、図 44 の高電圧演算増幅器 4403 に使用することができる高電圧演算増幅器 4600 を示す。高電圧演算増幅器 4600 は、示された配置で示される構成要素を備える。HV カスコードバイアスノード VCASN2A 及び VCASN2B は、トランジスタにわたる最大電圧ドロップを最小化するために、電圧値が出力電圧 VOUT に依存するように実装される。例えば、ノード電圧 VOUT が高いとき、VCASN1B 及び VCA

10

20

30

40

50

S2Bは、高い。

【0166】

図47は、図44の高電圧演算増幅器4403の適応高電圧カスコードバイアスを提供するために使用することができる適応高電圧電源4700を示す。適応高電圧電源4700は、示された配置で示される構成要素を備える。

【0167】

図48は、ビット線ドライバ3411の各々に使用することができる列ドライバ4800を示す。列ドライバ4800は、図示の構成では、ラッチ4801、インバータ4802、NORゲート4803、PMOSトランジスタ4804、NMOSトランジスタ4805及び4806、並びにセンス増幅器4807を備える。図示のように、VCASALEベルは、より高いレベルのVIN、すなわち、約 $V_{IN} + 2 \cdot V_{T_PMOS}$ を追跡する。VCASBLEベルは、より低いレベルのVIN、すなわち、約 $V_{IN} - V_{T_NMOS}$ を追跡する。異なるMOSトランジスタ及び異なる $I \cdot R$ 電圧ドロップ（電流経路への抵抗Rの挿入など）によって、他の値が可能である。

【0168】

図49は、図48のセンス増幅器4807に使用することができる、センス増幅器4900を示す。センス増幅器4900は、図示の構成では、調整可能な電流基準源4901、スイッチ4902、NMOSトランジスタ4903、コンデンサ4904、スイッチ4905、電流源4906、及びインバータ4907を備える。センス増幅器4907は、VMMアレイ3408内のメモリセル4908に結合されている。

【0169】

図50は、ビット線基準デコーダ5001及び基準セル5001₀～5002_Nを備える基準アレイ回路5000を示す。

【0170】

図51は、ビット線基準デコーダ5101及び基準セル5102₀～5100_Nを備える基準アレイ回路5100を示す。

精密プログラミング回路及びアルゴリズム

【0171】

図52は、演算増幅器5201、抵抗器5203、及び可変抵抗器5202（低電圧ドメイン可変抵抗器であり得る）を備える、HV演算増幅器4403用の電圧を提供するための適応高電圧電源5200を示す。適応高電圧電源5200は、入力VINを受け取り、高電圧信号HVOUTを生成し、ここでゲインは、抵抗トリム回路ネットワーク（図示せず）によって可変抵抗器5202の抵抗を調整することによって調整することができる。一実施形態では、抵抗器5203は、低電圧ドメイン（例えば、1V又は1.8V）にあり、低電圧デバイスを使用し、演算増幅器5201、及びトリム回路ネットワークを有する可変抵抗器5202は、高電圧ドメイン（例えば、12V）にあり、高電圧デバイスを使用する。次いで、HVOUTを使用して、不揮発性メモリセルをプログラムすることができる。

【0172】

図53は、演算増幅器5301、可変抵抗器5303、及び抵抗器5302を備える、HV演算増幅器3403用の電圧を提供するための適応高電圧電源5300を示す。適応高電圧電源5300は、入力VINを受け取り、高電圧信号HVOUTを生成し、ここでゲインは、抵抗トリム回路ネットワーク（図示せず）によって可変抵抗器5303の抵抗を調整することによって調整することができる。一実施形態では、可変抵抗器5303及びトリム回路ネットワークは、低電圧ドメイン（例えば、1V又は1.8V）にあり、低電圧デバイスを使用し、演算増幅器5301及び抵抗器5302は、高電圧ドメイン（例えば、12V）にあり、高電圧デバイスを使用する。次いで、HVOUTを使用して、不揮発性メモリセルをプログラムすることができる。

【0173】

図54は、演算増幅器5401、抵抗器5402、抵抗器5403、演算増幅器540

4、及び調整可能な分圧器5405を備える、HV演算増幅器4403用の電圧を提供するための適応高電圧電源5400を示す。調整可能な分圧器5405は、電圧源 V_S を受け取り、抵抗器5408と、iセットの抵抗器5406i及びスイッチ5407iと、を備える。演算増幅器5404の非反転端子上の入力電圧でもあり、入力電圧 V_{IN} である、調整可能な分圧器5405によって出力される電圧は、スイッチ5407iのうちのどれが閉じているかに応じて変動する。入力 V_{IN} にตอบสนองして、高電圧信号 HV_{OUT} が生成される。ここで、 V_{IN} の大きさは、調整可能な分圧器5405によって調整することができる。一実施形態では、調整可能な分圧器5405及び演算増幅器5404は、低電圧ドメイン（例えば、1V又は1.8V）であり、低電圧デバイスを使用し、演算増幅器5401は、高電圧ドメイン（例えば、12V）であり、高電圧デバイスを使用する。次いで、 HV_{OUT} を使用して、不揮発性メモリセルをプログラムすることができる。

【0174】

図55は、調整可能な分圧器5505及び微細抵抗器HVネットワーク5580を備える、HV演算増幅器4403用の電圧を提供するための適応高電圧電源5500を示す。微細抵抗器ネットワーク5580は、バッファ5501と、バッファ5502と、jセットの抵抗器5504j及びスイッチ5504jを備える調整可能な分圧器5503と、を備える。調整可能な分圧器5505は、電圧源 V_S を受け取り、抵抗器4408と、iセットの抵抗器5506i及びスイッチネットワーク5507iと、を備える。調整可能な分圧器5505によって出力される電圧は、スイッチネットワーク5507iのスイッチのうちのどれが閉じているかに応じて変動する。調整可能な分圧器5503は、高電圧 $HV_COARSE1$ 及び $HV_COARSE2$ を受け取る。高電圧 $HV_COARSE1$ は更に、適応高電圧電源5500の第1の出力である。 HV_FINE である調整可能な分圧器5503によって出力される電圧は、 HV_COARSE 、及びスイッチネットワーク5504jのスイッチのうちのどれが閉じているかに応じて変動する。ここで、 $HV_COARSE1/2$ の大きさは、調整可能な分圧器5505内で閉じられているスイッチネットワーク5507iのスイッチを変更することによって調整することができる。 HV_FINE の大きさは、調整可能な分圧器5503内で閉じられているスイッチネットワーク5504jのスイッチを変更することによって調整することができる。数値例として、調整可能な分圧器5505は、1ステップごとに200mV（すなわち、電圧インクリメント、1つの抵抗器5506にわたる電圧）を提供することができ、 $HV_COARSE1$ 及び $HV_COARSE2$ にわたる電圧は600mVであり、調整可能な分圧器5503は、1ステップごとに5mV（すなわち、電圧インクリメント、1つの抵抗器5504jにわたる電圧）を提供することができる。これらの高電圧を使用して、不揮発性メモリセルをプログラムすることができる。

【0175】

図56は、粗SC（スイッチキャップ）ネットワーク5650及び微細抵抗器HVネットワーク5680を備える、HV演算増幅器4403用の電圧を提供するための適応高電圧電源5600を示す。粗SCネットワーク5650は、演算増幅器5601、SCネットワーク5609、及びSCネットワーク5608を備える。SCネットワーク5609は、静電容量CFBの調整可能なコンデンサ5602を備える。SCネットワーク5608は、静電容量CINの調整可能なコンデンサ5603、並びにスイッチ5604、5605、5606、及び5607を備える。ここで、 $HV_{OUT} = V_{IN} * (C_{IN} / C_{FB})$ である。微細抵抗器HVネットワーク5680は、図55のネットワーク5580に類似している。粗SCネットワーク5650は、粗い調整可能なレベル、例えば200mVのステップを提供し、微細抵抗器HVネットワーク5680は、微細レベル、例えば5mVのステップを提供する。 HV_{OUT} を使用して、不揮発性メモリセルをプログラムすることができる。

【0176】

図52、43、52～56に示されるHV演算増幅器5403の適応HV電源は、出力電圧の関数として電力を最適化するために使用される。 $V_{HVO PA} = V_{OUT} + dV$ で

10

20

30

40

50

あり、例えば、 $V_{HVO PA} = 6V$ 、 $V_{OUT} = 4V$ 及び $dV = 2V$ である。基本的に、 $HVO PA4403$ は、最大HV電圧（例えば、 $12V$ ）を常に供給する必要はない。

【0177】

図57は、ハードウェア実装を簡略化する1つの単一レベル基準のみを使用した、修正SAR（Successive Approximation：逐次近似）シーケンシャル検証アルゴリズム5700を示す。図は、4つの入力デジタルビットに対して比較される4つの出力デジタルビットにセル出力を変換するために使用される4ビット検証アルゴリズムを示す。最大有効ビット $DO UT3$ は、中央基準値を基準線に適用し、セル出力を基準レベルと比較することによって最初に変換される。次の有効ビット $DO UT2$ は、中央基準値を上半分（すなわち、 $L8 \sim L15$ ）に適用し、セル出力を中央基準レベルと比較し、次いで、中央基準値を下半分（すなわち、 $L7 \sim L0$ ）に適用し、セル出力を中央基準レベルと比較することによって次に変換される。次のデジタルビットは同様に変換される。このアプローチは、4ビット出力のために15個の変換パルス（ステップ）を必要とする。セル出力は、メモリセル内の格納された重みから変換される電流又は電圧のいずれかである。

【0178】

図58は、変換パルス数が半分減少した2つの基準線を使用した、修正SARシーケンシャル検証アルゴリズム5800を示す。最大有効ビットは、単一の基準線を使用して上記のように行われる。次のシーケンシャル変換ステップは、2つの基準線を使用する。各基準線については、変換ステップは上記と同様である。このアプローチは、4ビット出力のために8つのステップを必要とする。

【0179】

上記の検証アルゴリズムを使用して、ニューロン電流（VMMアレイからの出力電流）をデジタルビットに変換することができる。

【0180】

図59は、不揮発性メモリセルのプログラム動作後の動作を検証する間に使用するための調整可能な2D温度計コード基準電流源5900を示す。2D温度計コード基準電流源5900は、バッファミラー5901（演算増幅器5902及びPMOSトランジスタ5903を備える）と、調整可能なバイアス源5904と、特定のデバイス5906がラベル5906-（行）（列）によって記されるデバイス5906の*i*行及び*j*列のアレイを備える2Dアレイ5905と、を備える。ここで、バッファミラー5901によって出力される基準電流の量を調整することができるように、デバイス5906の様々な組み合わせを活性化させることができる。図示のように、2Dアレイ5905内に16個の電流ミラー（デバイス5906）が存在する。調整可能な2D温度計コード基準電流源5900は、基本的に、4つのデジタル入力コードを、バイアス源5904から提供される $I_{bias unit}$ の1～16倍の値で基準電流バイアスに変換する。これらは、例えば、図58に示されるようなVMMアレイ内のメモリセルの16個のレベルに対応する。

【0181】

例えば、バイアス源5904は、デバイス5906にミラーリングされる $1nA$ の電流 $I_{bias unit}$ を提供することができる。ここで、第1の行は、デバイス5906-11～5906-1*j*からなり、一度に1つのデバイス5906で左から右に順次イネーブルされる。次に、次の行が左から右に順次イネーブルされて、第1の列に追加され、これは5つ、6つ、7つ、次いで8つのデバイス5906がイネーブルされることを意味する。したがって、デバイス5906を順次イネーブルすることによって、従来のバイナリデコードに関連するトランジスタ不整合状況を回避することができる。次いで、イネーブルされたデバイス5906の合計は、バッファミラー5901によってミラーリングされ、調整可能な電流として出力され、図39の調整可能な電流基準源3901に使用することができる。バイアス源5904は、 $50pA / 100pA / 200pA / \dots / 100nA$ などのトリミング可能な範囲のユニットバイアスを提供することができる。図示の調整可能な 4×4 の2D温度計コード基準電流源5900は、 32×32 又は 8×32 などの任意の他の寸法であり得る。

10

20

30

40

50

【0182】

図60は、図59のデバイス5906に使用することができる、基準サブ回路6000を示す。基準サブ回路6000は、図示されるように構成された、NMOSトランジスタ6001及び6002を備える。トランジスタ6002は、電流ミラーバイアストランジスタであり、トランジスタ6001は、(バイアストランジスタ6002が出力ノードOUTPUTに接続されることを可能にするための) イネーブルトランジスタである。

【0183】

本明細書で使用される場合、「の上方に(over)」及び「に(on)」という用語は両方とも、「の上に直接」(中間材料、要素、又は間隙がそれらの間に配設されていない)、及び「の上に間接的に」(中間材料、要素、又は間隙がそれらの間に配設されている)を包括的に含むことに留意されるべきである。同様に、「隣接した」という用語は、「直接隣接した」(中間材料、要素、又は間隙がそれらの間に配設されていない)、及び「間接的に隣接した」(中間材料、要素、又は間隙がそれらの間に配設されている)を含み、「に取り付けられた」は、「に直接取り付けられた」(中間材料、要素、又は間隙がそれらの間に配設されていない)、及び「に間接的に取り付けられた」(中間材料、要素、又は間隙がそれらの間に配設されている)を含み、「電氣的に結合された」は、「に直接電氣的に結合された」(要素と一緒に電氣的に接続する中間材料又は要素がそれらの間にない)、及び「に間接的に電氣的に結合された」(要素と一緒に電氣的に接続する中間材料又は要素がそれらの間にある)を含む。例えば、要素を「基板の上方に」形成することは、その要素を基板に直接、中間材料／要素をそれらの間に伴わずに形成すること、及びその要素を基板の上に間接的に1つ以上の中間材料／要素をそれらの間に伴って形成することを含み得る。

10

20

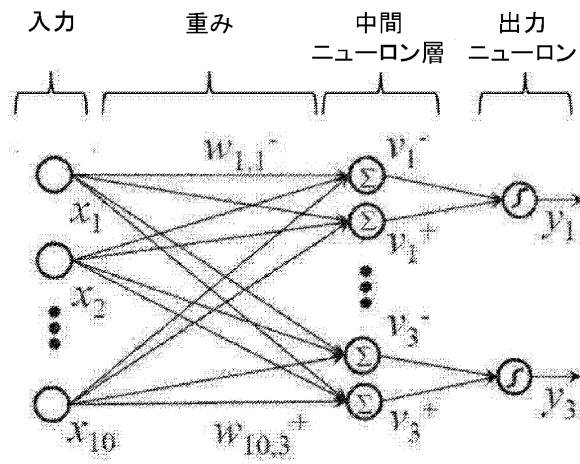
30

40

50

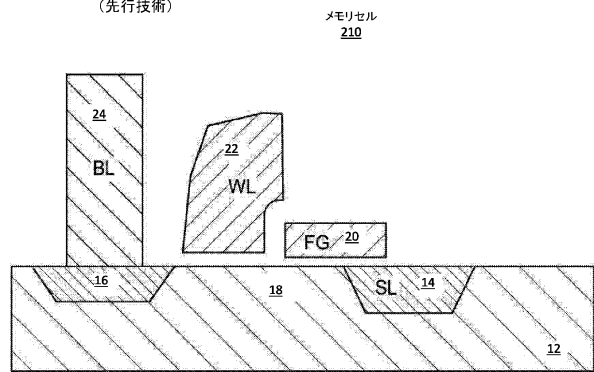
【図面】

【図 1】



【図 2】

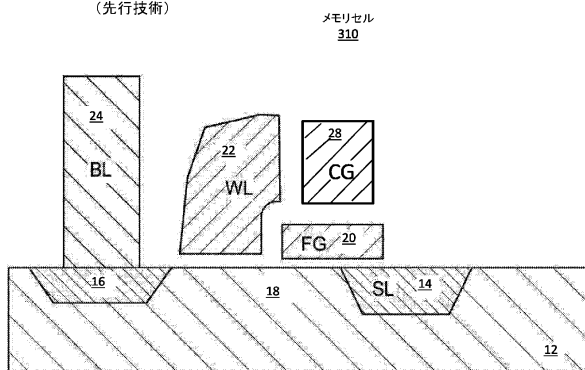
(先行技術)



(先行技術)

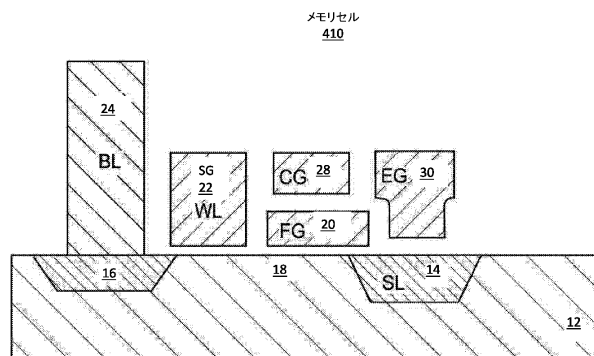
【図 3】

(先行技術)

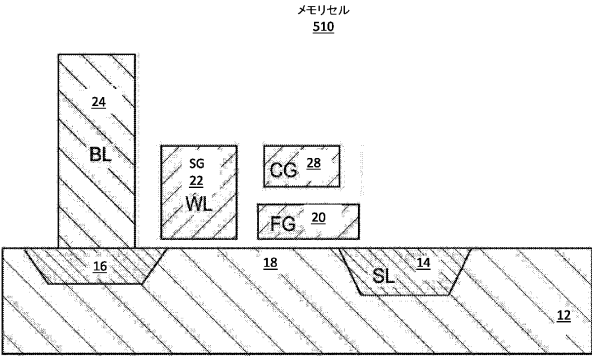


【図 4】

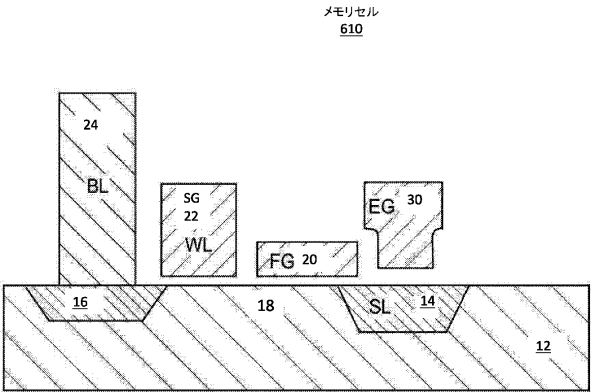
(先行技術)



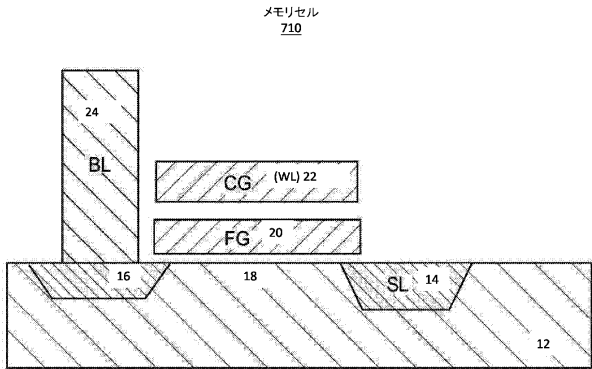
【図 5】
(先行技術)



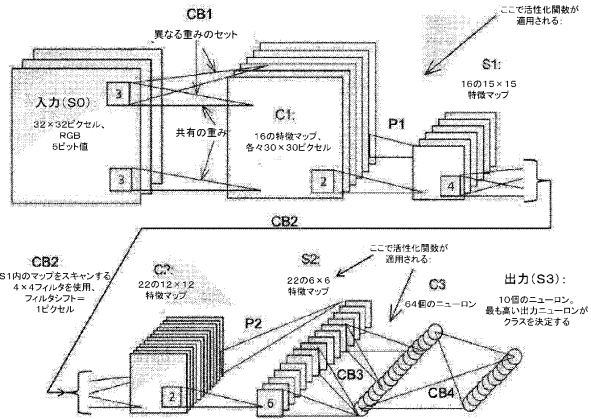
【図 6】
(先行技術)



【図 7】
(先行技術)



【図 8】



10

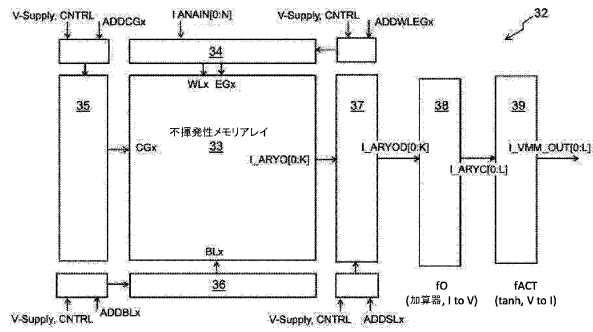
20

30

40

50

【図 9】



【図 10】

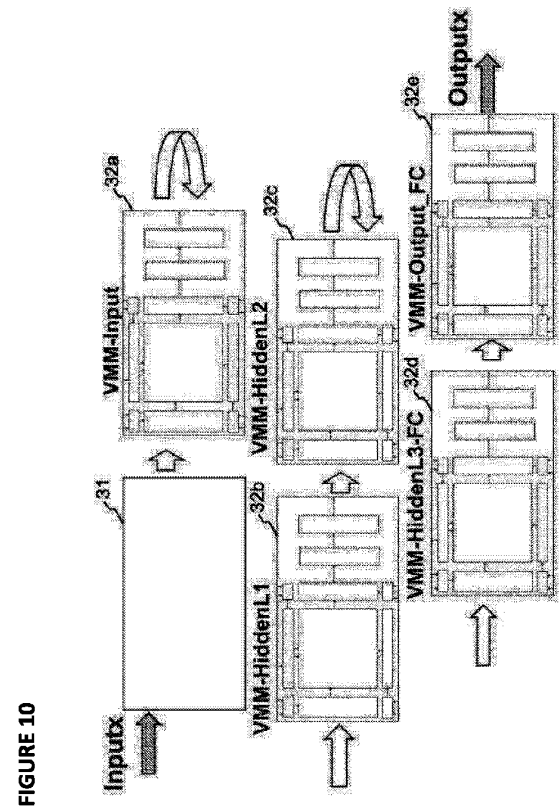
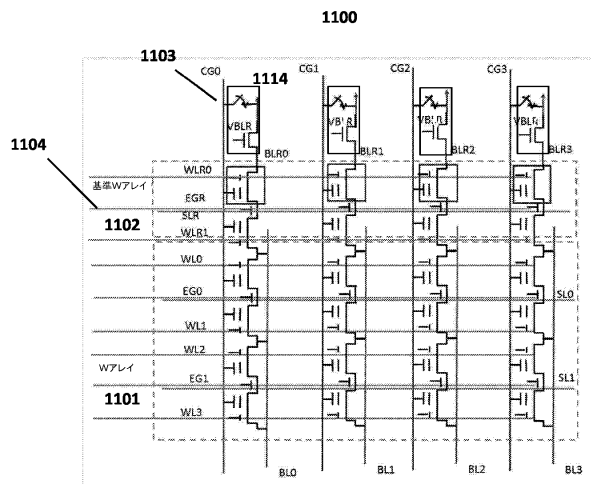
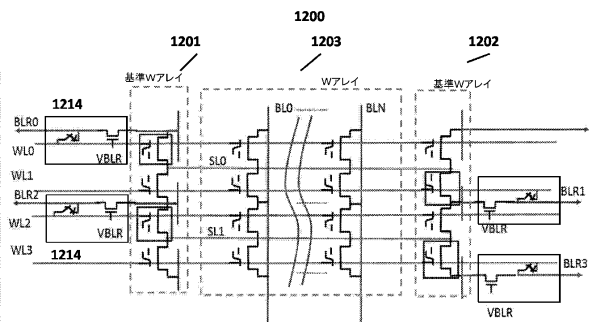


FIGURE 10

【図 11】



【図 12】



10

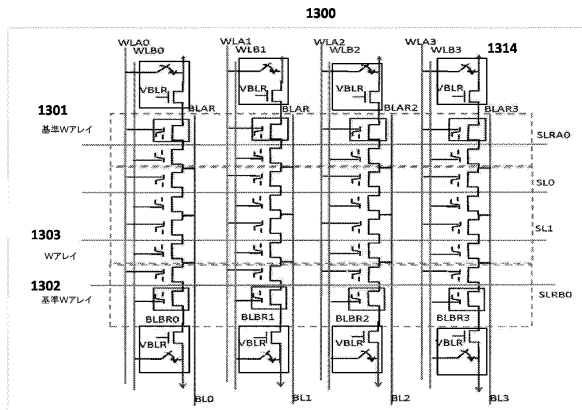
20

30

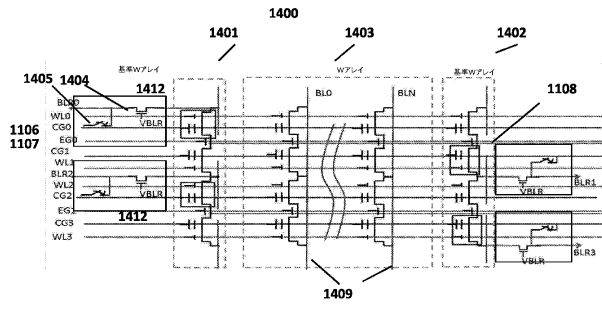
40

50

【図 13】

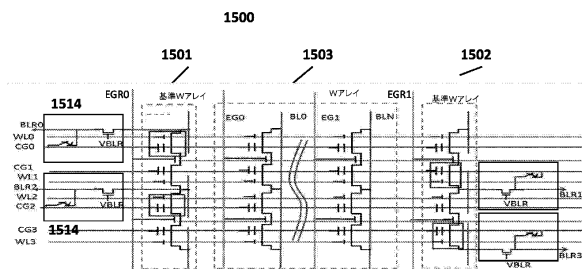


【図 14】



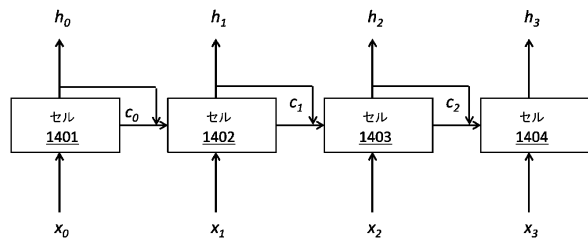
10

【図 15】



【図 16】

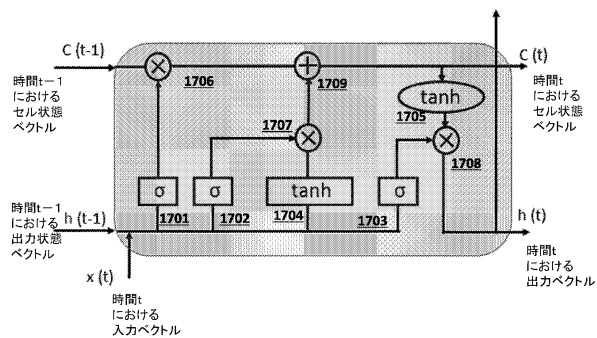
(先行技術)

LSTM
1600

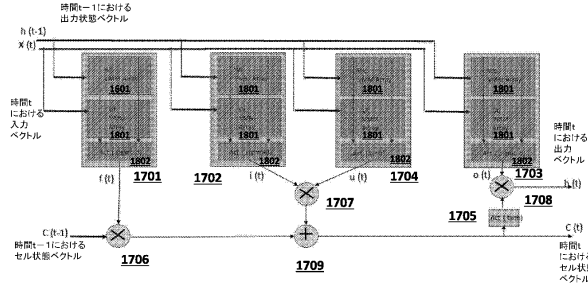
20

【図 17】

(先行技術)

LSTMセル
1700

【図 18】

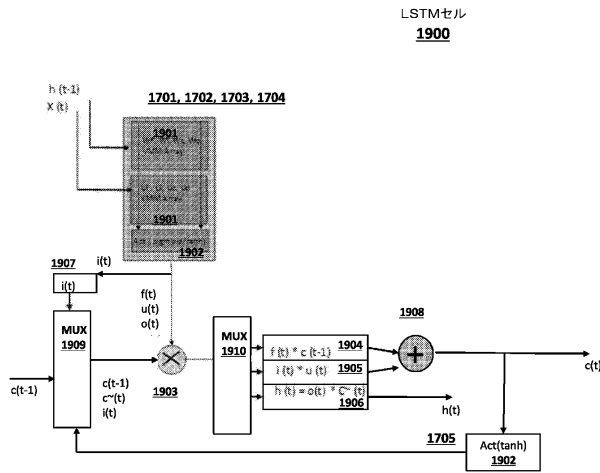
LSTMセル
1800

30

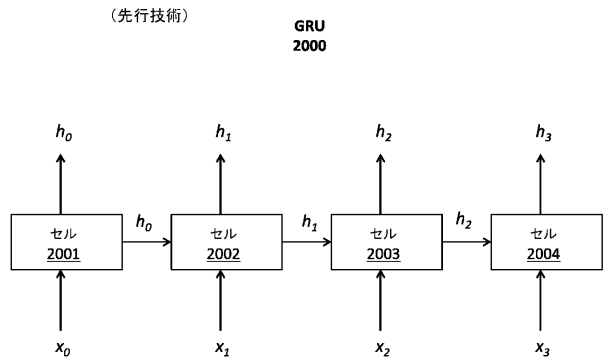
40

50

【図 19】

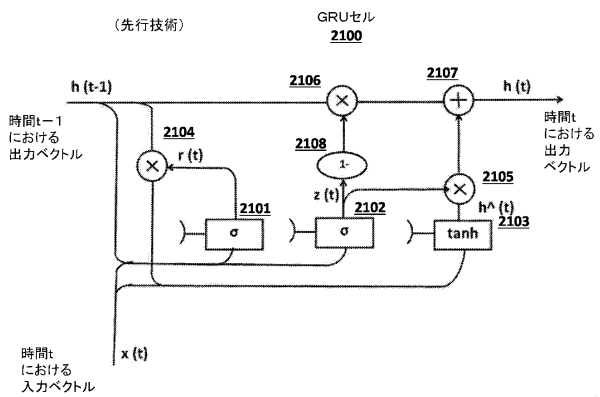


【図 20】

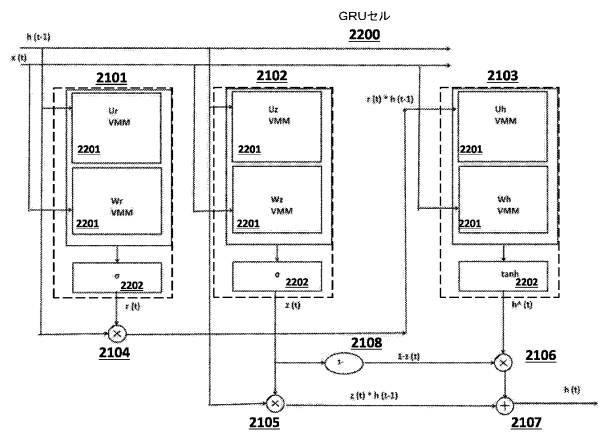


10

【図 21】



【図 22】



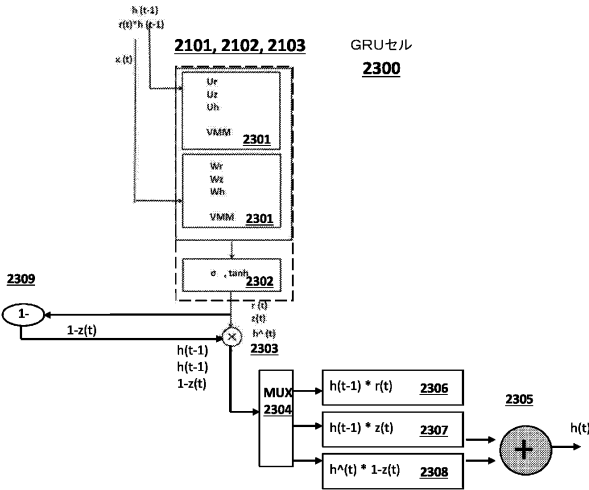
20

30

40

50

【図 2 3】



【図 2 4】

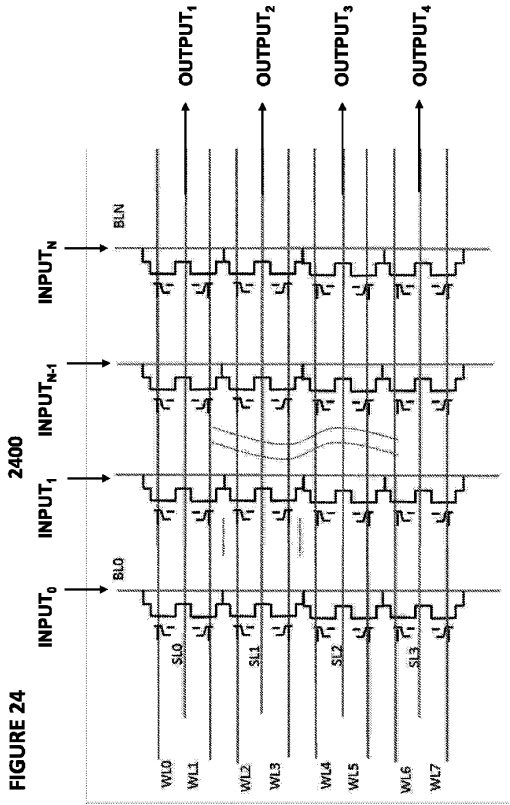


FIGURE 24

【図 2 5】

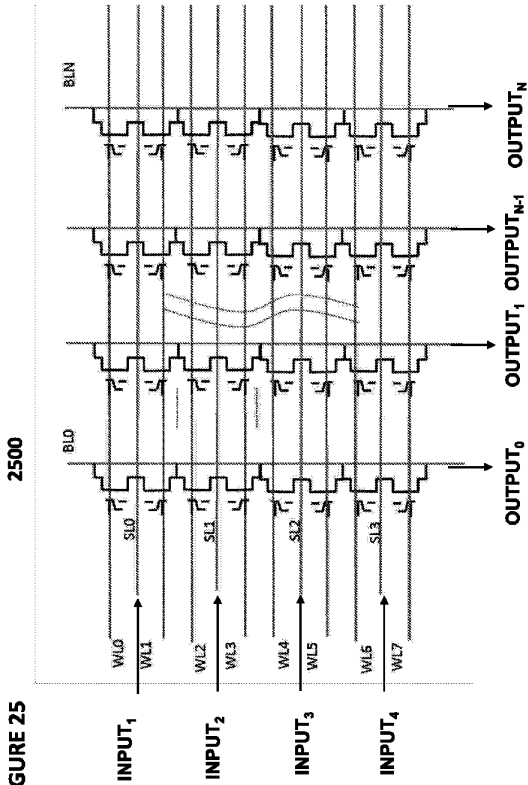


FIGURE 25

【図 2 6】

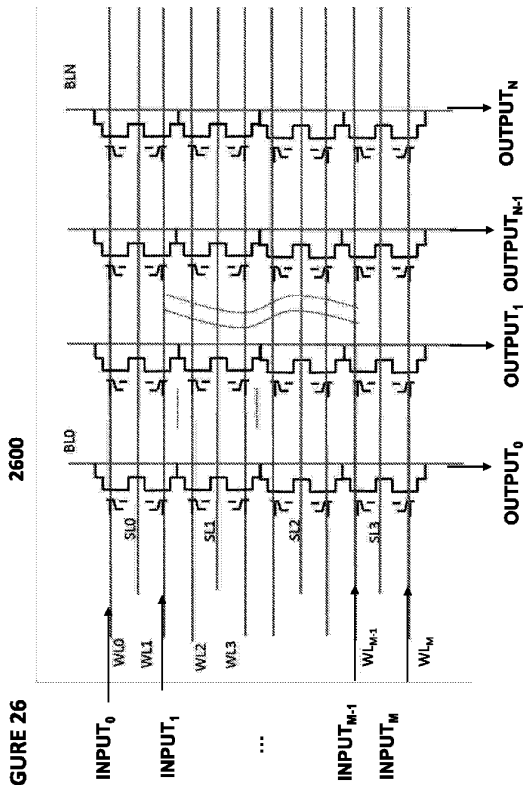


FIGURE 26

10

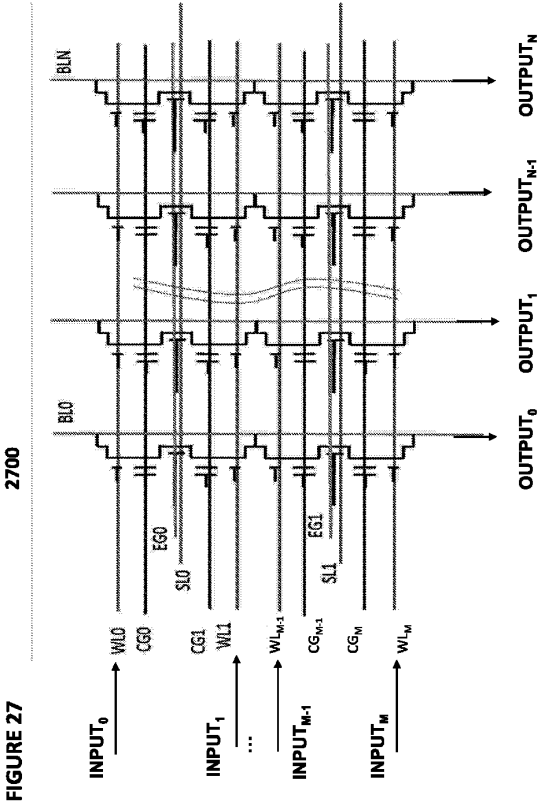
20

30

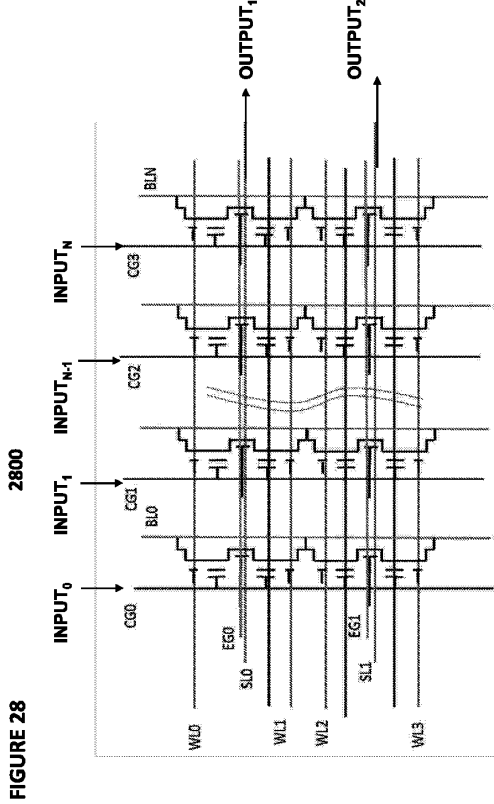
40

50

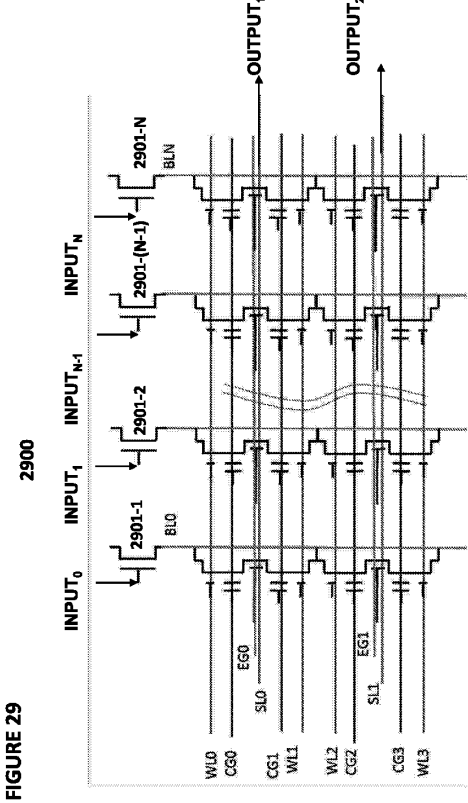
【図 27】



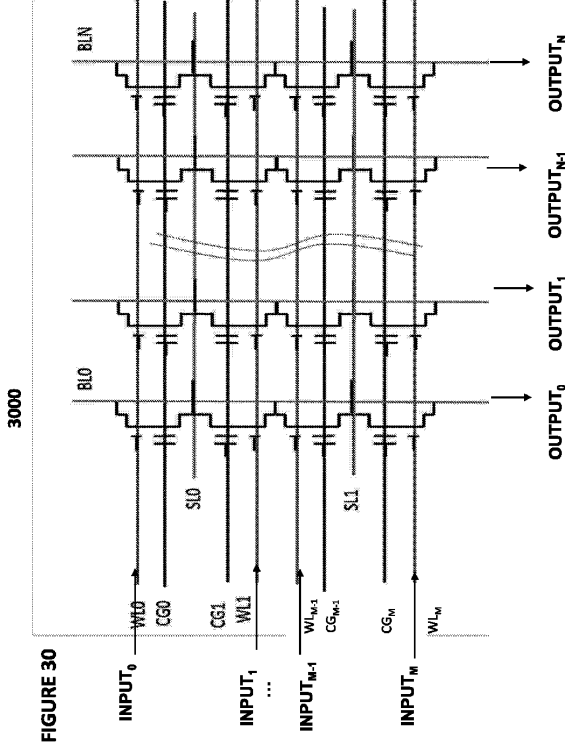
【図 28】



【図 29】



【図 30】



10

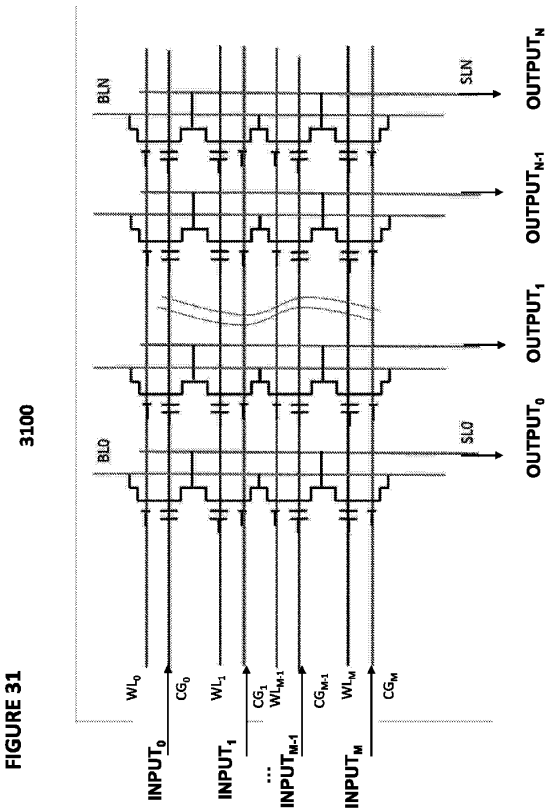
20

30

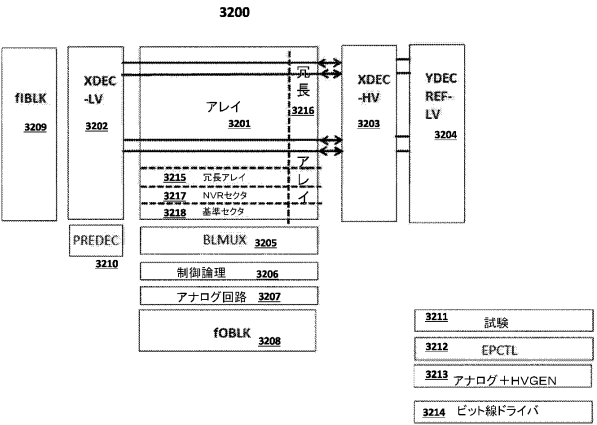
40

50

【図 3 1】



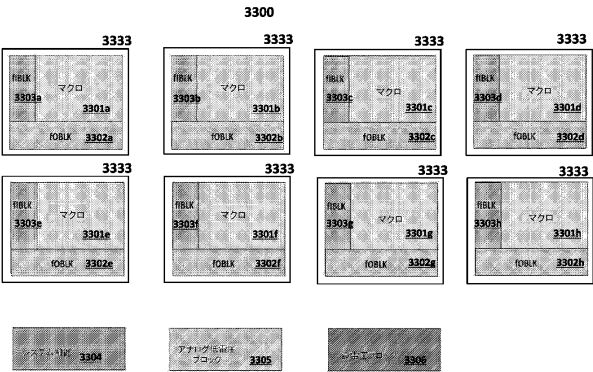
【図 3 2】



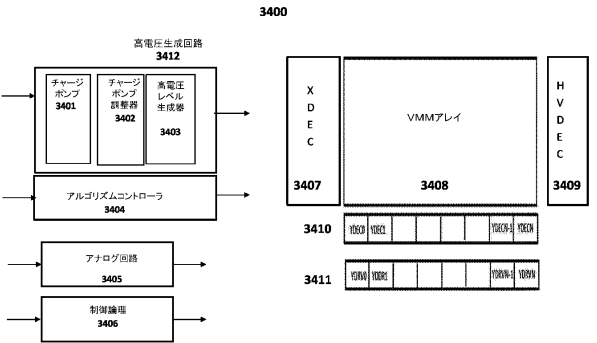
10

20

【図 3 3】



【図 3 4】

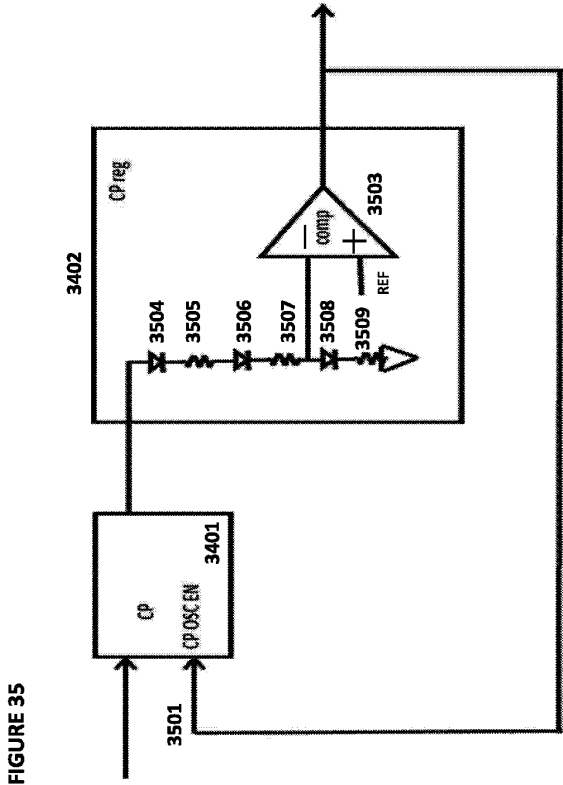


30

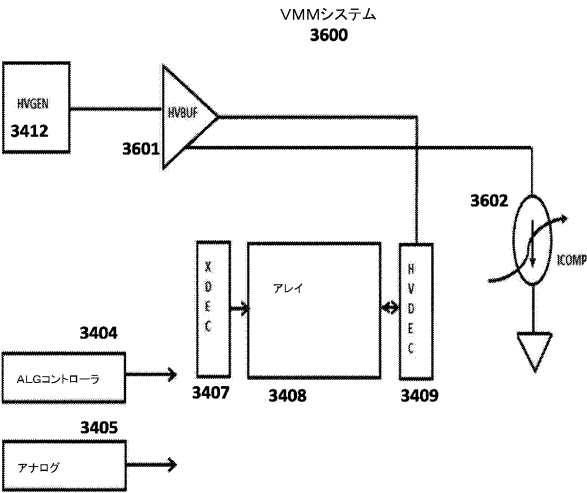
40

50

【図 35】



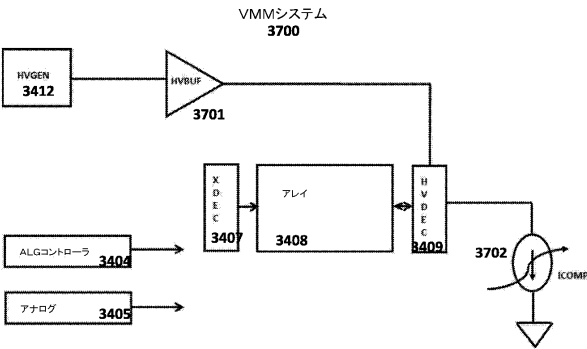
【図 36】



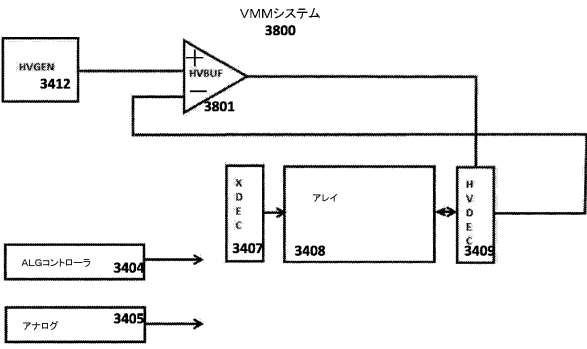
10

20

【図 37】



【図 38】



30

40

50

【図 40】

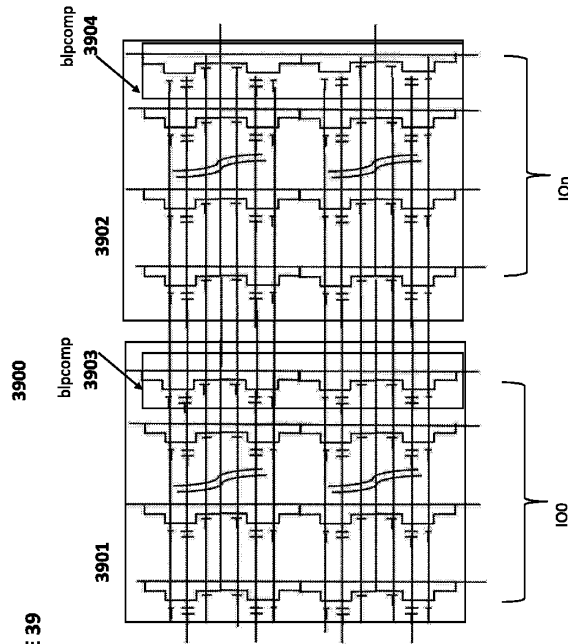


FIGURE 39

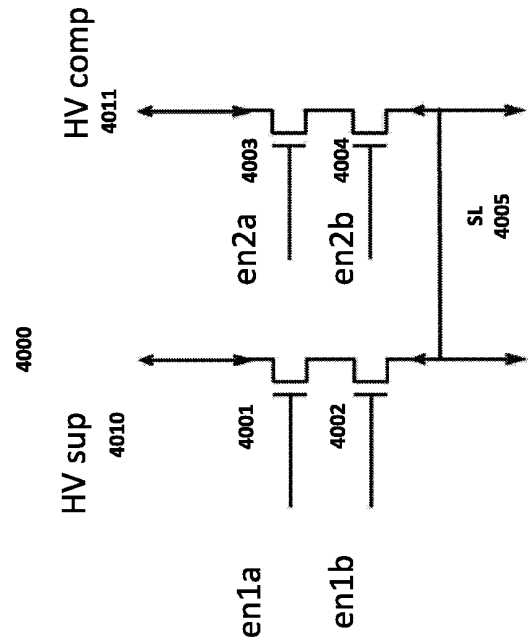
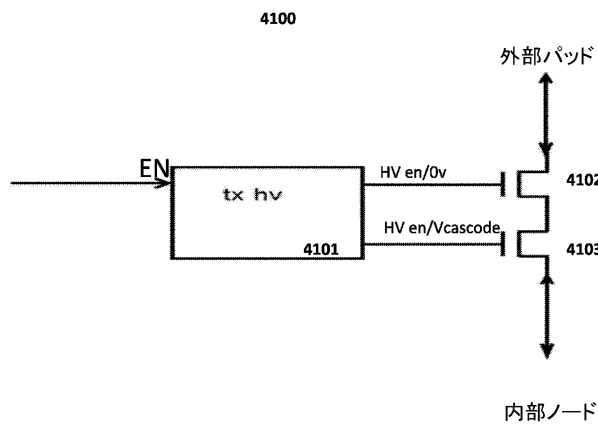
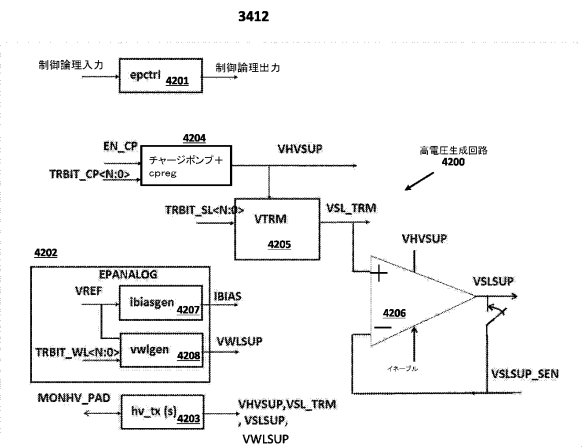


FIGURE 40

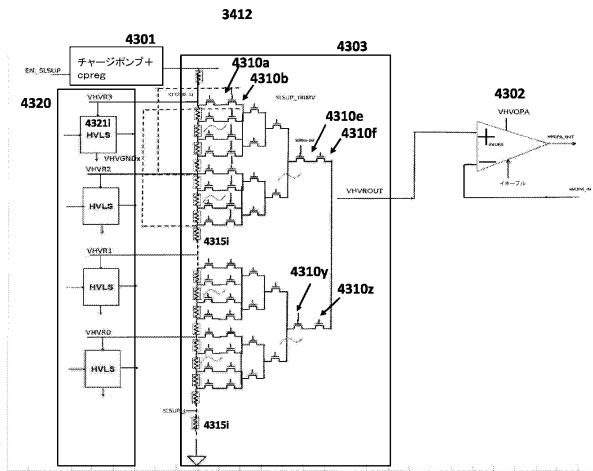
【図 4 1】



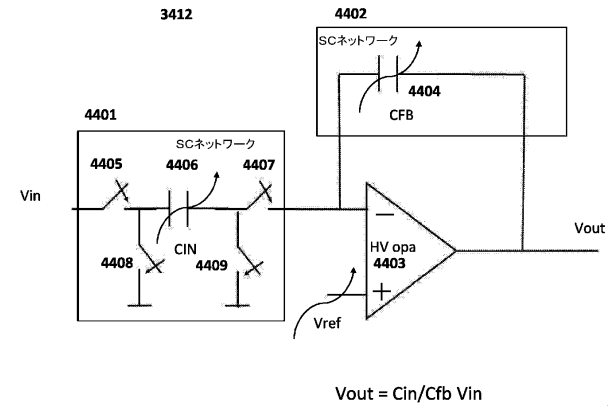
【図 4 2】



【図 4 3】

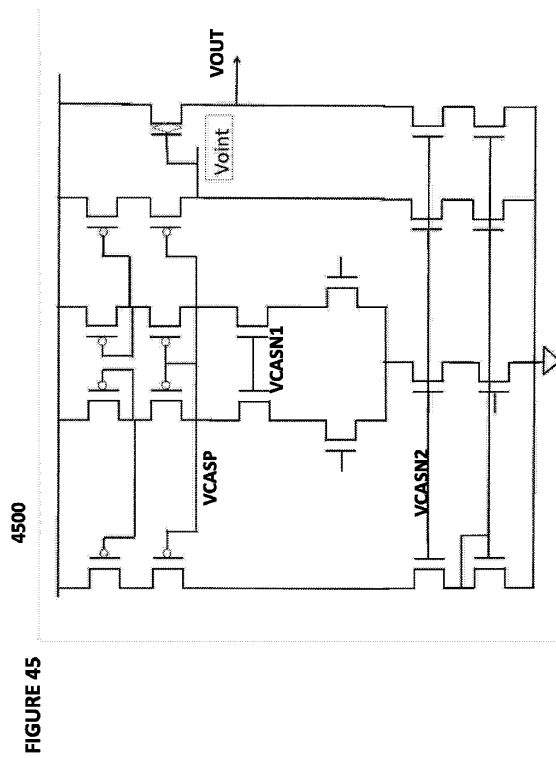


【図 4 4】

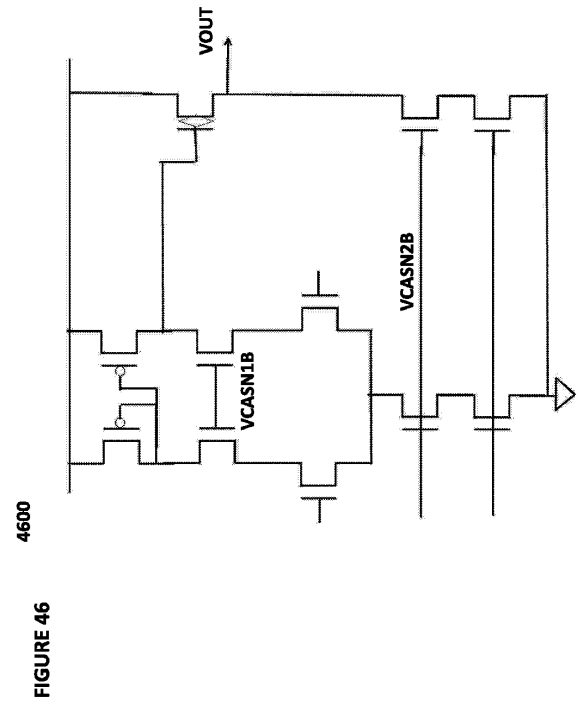


10

【図 45】



【図 46】



20

30

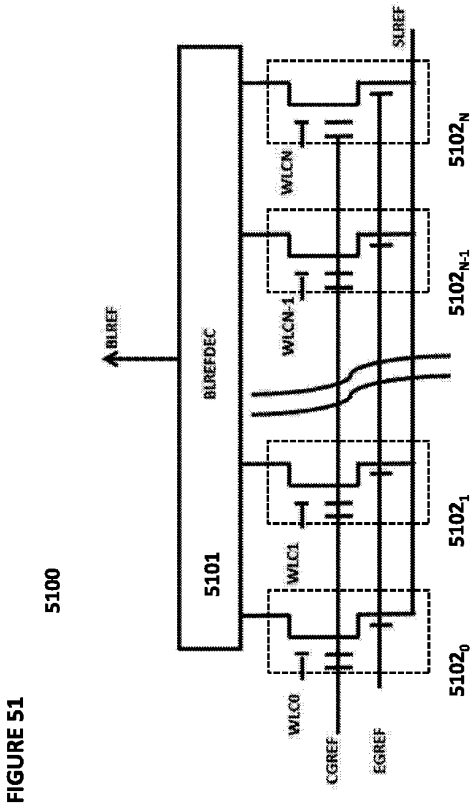
40

50

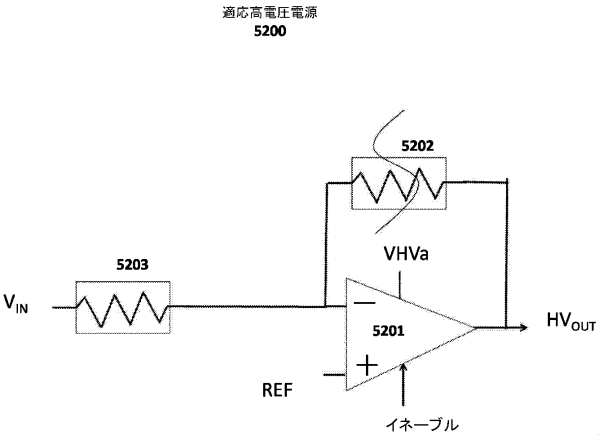
[illegible]

FIGURE 50

【図 5 1】



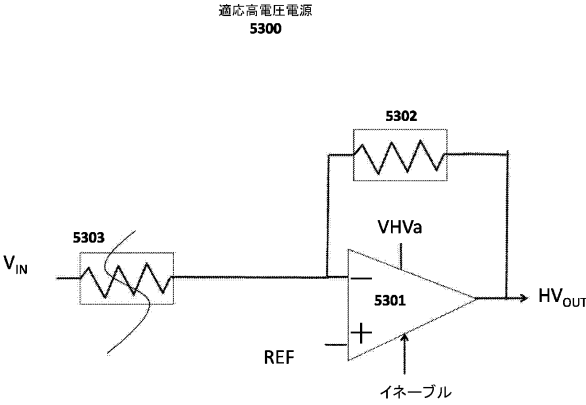
【図 5 2】



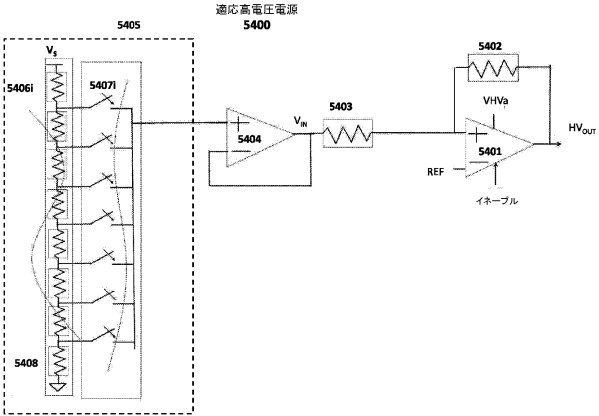
10

20

【図 5 3】



【図 5 4】

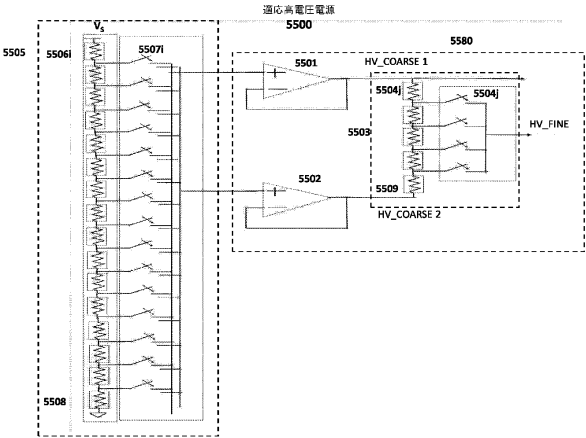


30

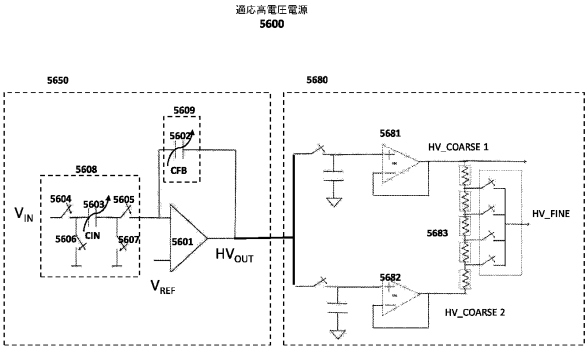
40

50

【図 5 5】



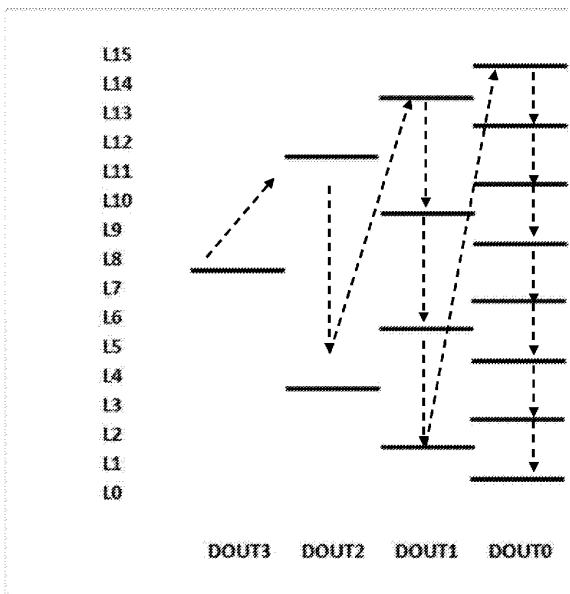
【図 5 6】



10

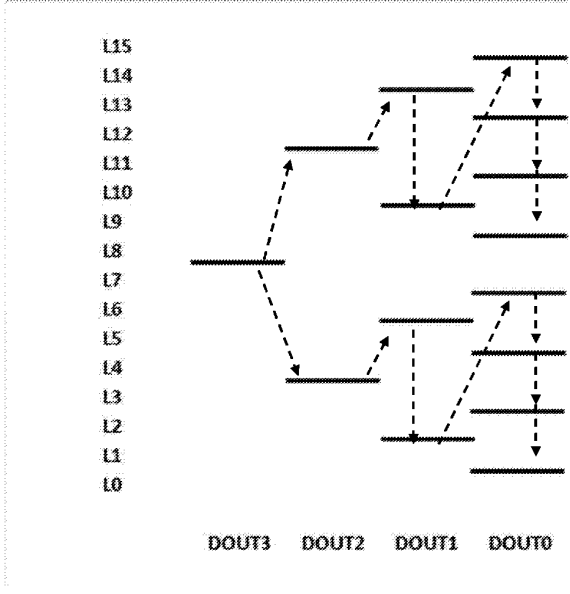
【図 5 7】

単一基準検証アルゴリズム
5700



【図 5 8】

二重基準検証アルゴリズム
5800



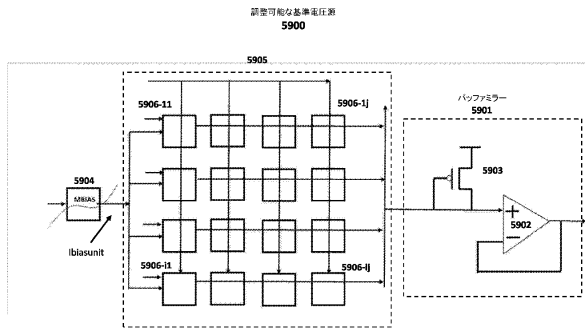
20

30

40

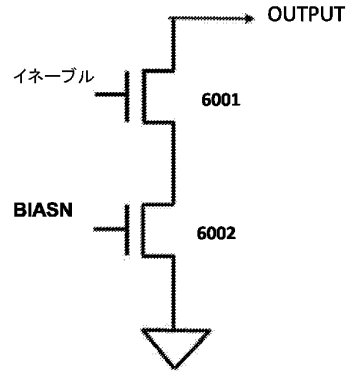
50

【図 59】



【図 60】

基準サブ回路
6000



10

20

30

40

50

フロントページの続き

(33)優先権主張国・地域又は機関

米国(US)

ニア州、サンノゼ、ダンナ コート 431

(72)発明者 トリン、ステファン

アメリカ合衆国 95148 カリフォルニア州、サンノゼ、チエリ ブレイス 3430

(72)発明者 リ、アン

アメリカ合衆国 95121 カリフォルニア州、サンノゼ、リンドミュア ドライブ 3385

(72)発明者 ホン、スタンレー

アメリカ合衆国 95131 カリフォルニア州、サンノゼ、プリストル ベイ コモン 1848

審査官 後藤 彰

(56)参考文献 特開2003-223791 (JP, A)

米国特許出願公開第2008/0056035 (US, A1)

韓国公開特許第2003-0002732 (KR, A)

国際公開第2017/200883 (WO, A1)

特表2019-517138 (JP, A)

(58)調査した分野 (Int.Cl., DB名)

G11C 11/54

G11C 11/56

G11C 16/08