(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6785738号
(P6785738)
(45)発行日　令和2年11月18日(2020.11.18)　(24)登録日　令和2年10月29日(2020.10.29)

(51)Int.Cl.　F I
G06F 12/00 (2006.01)　G06F 12/00 560F

請求項の数 9　(全 23 頁)

(21)出願番号　特願2017-193447(P2017-193447)
(22)出願日　平成29年10月3日(2017.10.3)
(65)公開番号　特開2018-73452(P2018-73452A)
(43)公開日　平成30年5月10日(2018.5.10)
審査請求日　令和2年9月9日(2020.9.9)
(31)優先権主張番号　62/414,426
(32)優先日　平成28年10月28日(2016.10.28)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　15/425,996
(32)優先日　平成29年2月6日(2017.2.6)
(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

(73)特許権者　390019839
三星電子株式会社
Ｓａｍｓｕｎｇ　Ｅｌｅｃｔｒｏｎｉｃｓ　Ｃｏ．，Ｌｔｄ．
大韓民国京畿道水原市霊通区三星路１２９　１２９，Ｓａｍｓｕｎｇ－ｒｏ，Ｙｅｏｎｇｔｏｎｇ－ｇｕ，Ｓｕｗｏｎ－ｓｉ，Ｇｙｅｏｎｇｇｉ－ｄｏ，Ｒｅｐｕｂｌｉｃ　ｏｆ　Ｋｏｒｅａ
(74)代理人　110000051
特許業務法人共生国際特許事務所
(72)発明者　李　双　辰
アメリカ合衆国，　９３１１７，　カリフォルニア州，　ゴレタ，　コレヒオ　ロード，　アパート　＃１３０２，　６５１０
最終頁に続く

(54)【発明の名称】ＤＲＡＭ基盤のプロセシングユニット

(57)【特許請求の範囲】
【請求項１】
ＤＰＵ（Ｄｙｎａｍｉｃ　Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）であって、
少なくとも３以上である第１の所定の数のカラム及び３以上である第２の所定の数のローを有するアレイ内に配置された複数のＤＲＡＭ基盤のコンピューティングセルを含む少なくとも１つのコンピューティングセルアレイを備え、
前記カラムの各々は、前記カラムの第１ロー及び第２ロー上で動作するロジック機能を提供し、前記カラムの第３ロー内に前記ロジック機能の結果を格納するように構成され、
前記少なくとも１つのコンピューティングセルアレイは、前記第１の所定の数の２倍である第３の所定の数のシフトラインを更に含み、
前記シフトラインの各々は、少なくとも１つの対応する第１のトランジスタを介して前記コンピューティングセルのカラムに連結され、
前記シフトライン及び前記対応する第１のトランジスタは、前記少なくとも１つのコンピューティングセルアレイ内において、選択されたカラムの２つのローのコンピューティングセルの内容を少なくとも２つのカラムで右方向又は左方向にシフトするように構成されることを特徴とするＤＰＵ。
【請求項２】
前記第１の所定の数のカラム内に配置された少なくとも１つのＤＲＡＭ基盤のメモリセルを含む少なくとも１つのデータセルアレイと、

前記コンピューティングセルのカラムの各々に連結されたセンスアンプと、を更に含み、
前記少なくとも１つのデータセルアレイのＤＲＡＭ基盤のメモリセルのカラムの各々は、対応するコンピューティングセルアレイのカラムに対応し、
前記センスアンプは、前記コンピューティングセルのカラムの読出しビットラインに電気的に連結される入力、及び前記コンピューティングセルのカラムの書込みビットラインに電気的に連結される出力を含むことを特徴とする請求項１に記載のＤＰＵ。
【請求項３】
ＤＰＵ（Ｄｙｎａｍｉｃ　Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）であって、
少なくとも３以上である第１の所定の数のカラム及び３以上である第２の所定の数のローを有するアレイ内に配置された複数のＤＲＡＭ基盤のコンピューティングセルを含む少なくとも１つのコンピューティングセルアレイと、
前記第１の所定の数のカラム及び少なくとも１つのロー内に配置された少なくとも１つのＤＲＡＭ基盤のメモリセルを含む少なくとも１つのデータセルアレイと、を備え、
前記カラムの各々は、前記カラムの第１ロー及び第２ロー上で動作するロジック機能を提供し、前記カラムの第３ロー内に前記ロジック機能の結果を格納するように構成され、
前記少なくとも１つのコンピューティングセルアレイは、前記第１の所定の数の２倍である第３の所定の数のシフトラインを更に含み、
前記シフトラインの各々は、少なくとも１つの対応する第１のトランジスタを介して前記コンピューティングセルのカラムに連結され、
前記シフトライン及び前記対応する第１のトランジスタは、前記少なくとも１つのコンピューティングセルアレイ内において、選択されたカラムの２つのローのコンピューティングセルの内容を少なくとも２つのカラムで右方向又は左方向にシフトするように構成され、
前記少なくとも１つのデータセルアレイのＤＲＡＭ基盤のメモリセルのカラムの各々は、対応するコンピューティングセルアレイのカラムに対応することを特徴とするＤＰＵ。
【請求項４】
ＤＰＵ（Ｄｙｎａｍｉｃ　Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）であって、
３以上である第１の所定の数のカラム及び少なくとも１つのロー内に配置された少なくとも１つのＤＲＡＭ基盤のメモリセルを含む少なくとも１つのデータセルアレイと、
複数のＤＲＡＭ基盤のコンピューティングセルを含む少なくとも１つのコンピューティングセルアレイと、
前記コンピューティングセルのカラムの各々に連結されたセンスアンプと、
前記コンピューティングセルの各々に電気的に連結されるデコーダーと、を備え、
前記コンピューティングセルアレイの各々は、前記データセルアレイの各々に対応し、前記第１の所定の数のカラム及び３以上である第２の所定の数のローを有するアレイ内に配置され、
前記コンピューティングセルのカラムの各々は、前記コンピューティングセルのカラムの第１ロー及び第２ロー上で動作するロジック機能を提供し、前記コンピューティングセルのカラムの第３ロー内に前記ロジック機能の結果を格納するように構成され、
前記少なくとも１つのコンピューティングセルアレイは、前記第１の所定の数の２倍である第３の所定の数のシフトラインを更に含み、
前記シフトラインの各々は、少なくとも１つの対応する第１のトランジスタを介して前記コンピューティングセルのカラムに連結され、
前記シフトライン及び前記対応する第１のトランジスタは、前記少なくとも１つのコンピューティングセルアレイ内において、選択されたカラムの２つのローのコンピューティングセルの内容を少なくとも２つのカラムで右方向又は左方向にシフトするように構成され、

前記センスアンプの各々は、前記コンピューティングセルのカラム内の前記コンピューティングセルのカラムの読出しビットラインに電気的に連結される入力、及び前記コンピューティングセルのカラムの前記コンピューティングセルのカラムの書込みビットラインに電気的に連結される出力を含み、

前記デコーダーは、前記カラムのコンピューティングセルを選択するための命令に対応するＤＲＡＭ基盤のアドレス信号を受信して前記カラムの第１ロー及び第２ロー上で前記ロジック機能を生成し、前記カラムの第３ロー内に前記ロジック機能の結果を格納することを特徴とするＤＰＵ。

【請求項５】

少なくとも１つのカラムの前記ＤＲＡＭ基盤のコンピューティングセルの各々は、３つのトランジスタ及び１つのキャパシターを含むＤＲＡＭメモリセルを含むことを特徴とする請求項１、３、又は４に記載のＤＰＵ。

【請求項６】

少なくとも１つのカラムの前記ＤＲＡＭ基盤のコンピューティングセルは、ＮＯＲロジック機能を提供することを特徴とする請求項５に記載のＤＰＵ。

【請求項７】

少なくとも１つのカラムの前記ＤＲＡＭ基盤のコンピューティングセルの各々は、１つのトランジスタ及び１つのキャパシターを含むＤＲＡＭメモリセルを含むことを特徴とする請求項１、３、又は４に記載のＤＰＵ。

【請求項８】

前記ＤＲＡＭ基盤のコンピューティングセルの各々は、前記ＤＲＡＭ基盤のコンピューティングセルのビットラインに連結されたＡＬＵ（Ａｒｉｔｈｍｅｔｉｃ　Ｌｏｇｉｃ　Ｕｎｉｔ）を更に含み、

前記ＡＬＵは、前記ロジック機能を提供することを特徴とする請求項７に記載のＤＰＵ。

【請求項９】

前記ＡＬＵは、ＮＯＲロジック機能を提供することを特徴とする請求項８に記載のＤＰＵ。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、メモリシステムに関し、より詳細には、ＤＲＡＭ基盤のプロセシングユニットに関する。

【背景技術】

【０００２】

一般的に、ＧＰＵ（Ｇｒａｐｈｉｃｓ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）及びＴＰＵ（Ｔｅｎｓｏｒ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）がディープラーニング（ｄｅｅｐ　ｌｅａｒｎｉｎｇ）プロセシングのために使用される。ディープラーニングプロセシングはＧＰＵ又はＴＰＵによって効率的に処理できない高度に並列化されたプロセシングを含む。

【先行技術文献】

【特許文献】

【０００３】

【特許文献１】米国特許第５，９０１，０９５号明細書
【特許文献２】米国特許第６，０３５，３８４号明細書
【特許文献３】米国特許第６，１９５，７３８号明細書
【特許文献４】米国特許第７，２９９，０９９号明細書
【特許文献５】米国特許第８，０４２，０８２号明細書
【特許文献６】米国特許第９，１３６，８７２号明細書

【特許文献７】米国特許第９，１９７，２８５号明細書
【特許文献８】米国特許第９，３１７，４８２号明細書
【特許文献９】米国特許第９，３７８，１８１号明細書
【特許文献１０】米国特許出願公開第２００７／０２２６５６７号明細書
【特許文献１１】米国特許出願公開第２０１１／０３０２３６６号明細書
【特許文献１２】米国特許出願公開第２０１２／０２４６３８０号明細書
【特許文献１３】米国特許出願公開第２０１５／００８９１６６号明細書
【特許文献１４】米国特許出願公開第２０１５／０１３１３８３号明細書
【特許文献１５】米国特許出願公開第２０１６／０１４７６６７号明細書
【特許文献１６】米国特許出願公開第２０１６／０１７３１０２号明細書
【特許文献１７】欧州特許出願公開第１１９３５０２号明細書
【特許文献１８】欧州特許出願公開第２５２３３５２号明細書
【非特許文献】
【０００４】
【非特許文献１】ＭＡＴＡＭ，Ｋｉｒａｎ ｅｔ ａｌ．，“Ｅｎｅｒｇｙ－Ｅｆｆｉｃｉｅｎｔ Ｌａｒｇｅ－Ｓｃａｌｅ Ｍａｔｒｉｘ Ｍｕｌｔｉｐｌｉｃａｔｉｏｎ ｏｎ ＦＰＧＡｓ”，２０１３ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｒｅｃｏｎｆｉｇｕｒａｂｌｅ Ｃｏｍｐｕｔｉｎｇ ａｎｄ ＦＰＧＡｓ（ＲｅＣｏｎＦｉｇ），Ｄｅｃｅｍｂｅｒ９－１１，２０１３（８ｐａｇｅｓ）．
【非特許文献２】ＮＯＤＡ，Ｈｉｄｅｙｕｋｉ ｅｔ ａｌ．，“Ａ ｃｏｓｔ－ｅｆｆｉｃｉｅｎｔ ｈｉｇｈ－ｐｅｒｆｏｒｍａｎｃｅ ｄｙｎａｍｉｃ ＴＣＡＭ ｗｉｔｈ ｐｉｐｅｌｉｎｅｄ ｈｉｅｒａｒｃｈｉｃａｌ ｓｅａｒｃｈｉｎｇ ａｎｄ ｓｈｉｆｔ ｒｅｄｕｎｄａｎｃｙ ａｒｃｈｉｔｅｃｔｕｒｅ”，ＩＥＥＥ Ｊｏｕｒｎａｌ ｏｆ Ｓｏｌｉｄ－Ｓｔａｔｅ Ｃｉｒｃｕｉｔｓ，Ｖｏｌ．４０，Ｎｏ．１，Ｊａｎｕａｒｙ３，２００５（１０ｐａｇｅｓ）．
【非特許文献３】ＳＥＳＨＡＤＲＩ，Ｖｉｖｅｋ ｅｔ ａｌ．，“Ｆａｓｔ Ｂｕｌｋ Ｂｉｔｗｉｓｅ ＡＮＤ ａｎｄ ＯＲ ｉｎ ＤＲＡＭ”，ＩＥＥＥ Ｃｏｍｐｕｔｅｒ Ａｒｃｈｉｔｅｃｔｕｒｅ Ｌｅｔｔｅｒｓ，Ｖｏｌ．１４，Ｎｏ．２，Ｍａｙ１８，２０１５（６ｐａｇｅｓ）．
【非特許文献４】ＳＥＳＨＡＤＲＩ，Ｖｉｖｅｋ ｅｔ ａｌ．，“ＲｏｗＣｌｏｎｅ：Ｆｆａｓｔ ａｎｄ Ｅｎｅｒｇｙ－Ｅｆｆｉｃｉｅｎｔ Ｉｎ－ＤＲＡＭ Ｂｕｌｋ Ｄａｔａ Ｃｏｐｙ ａｎｄ Ｉｎｉｔｉａｌｉｚａｔｉｏｎ”，ＭＩＣＲＯ－４６，Ｐｒｏｃ．ｏｆ ４６ｔｈ ＩＥＥＥ／ＡＣＭ Ｉｎｔ．Ｓｙｍｐ．ｏｎ Ｍｉｃｒｏａｒｃｈｉｔｅｃｔｕｒｅ，Ｄｅｃｅｍｂｅｒ７，２０１３，ｐｐ．１８５－１９７．
【非特許文献５】ＷＡＮＧ，Ｇｅｓａｎ ｅｔ ａｌ．，“ＴＣＡＭ－Ｂａｓｅｄ Ｆｏｒｗａｒｄｉｎｇ Ｅｎｇｉｎｅ ｗｉｔｈ Ｍｉｎｉｍｕｍ Ｉｎｄｅｐｅｎｄｅｎｔ Ｐｒｅｆｉｘ Ｓｅｔ（ＭＩＰＳ） ｆｏｒ Ｆａｓｔ Ｕｐｄａｔｉｎｇ”，２００６ ＩＥＥＥ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｃｏｍｍｕｎｉｃａｔｉｏｎｓ，Ｊｕｎｅ１１－１５，２００６（７ｐａｇｅｓ）．
【発明の概要】
【発明が解決しようとする課題】
【０００５】
本発明は、上記従来技術に鑑みてなされたものであって、本発明の目的は、ＤＲＡＭ基盤のプロセシングユニットを提供することにある。
【課題を解決するための手段】
【０００６】
上記目的を達成するためになされた本発明の一態様によるＤＰＵ（Ｄｙｎａｍｉｃ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）は、少なくとも１つのカラムを含むアレイ内に配置された複数のＤＲＡＭ基盤のコンピューティングセルを含む少なくとも１つのコンピューティングセルアレイを備え、前記少なくとも１

つのカラムは、少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルを含み、前記少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルは、前記少なくとも３つのローの第１ロー及び第２ロー上で動作するロジック機能を提供し、前記少なくとも３つのローの第３ロー内に前記ロジック機能の結果を格納する。

【０００７】

上記目的を達成するためになされた本発明の他の態様によるＤＰＵ（Ｄｙｎａｍｉｃ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）は、少なくとも１つのカラムを含むアレイ内に配置された複数のＤＲＡＭ基盤のコンピューティングセルを含む少なくとも１つのコンピューティングセルアレイと、少なくとも１つのカラム内に配置された少なくとも１つのＤＲＡＭ基盤のメモリセルを含む少なくとも１つのデータセルアレイと、を備え、前記少なくとも１つのカラムは、少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルを含み、前記少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルは、前記少なくとも３つのローの第１ロー及び第２ロー上で動作するロジック機能を提供し、前記少なくとも３つのローの第３ロー内に前記ロジック機能の結果を格納する。

【０００８】

上記目的を達成するためになされた本発明の更に他の態様によるＤＰＵ（Ｄｙｎａｍｉｃ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）は、少なくとも１つのカラム内に配置された少なくとも１つのＤＲＡＭ基盤のメモリセルを含む少なくとも１つのデータセルアレイと、少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルを含む少なくとも１つのカラムを含むアレイ内に配置された複数のＤＲＡＭ基盤のコンピューティングセルを含む少なくとも１つのコンピューティングセルアレイと、前記少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルの読出しビットラインに電気的に連結された入力、及び前記少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルの書込みビットラインに電気的に連結された出力を含むセンスアンプと、前記少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルに電気的に連結されたデコーダーと、を備え、前記少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルは、前記少なくとも３つのローの第１ロー及び第２ロー上で動作するロジック機能を提供し、前記少なくとも３つのローの第３ロー内に前記ロジック機能の結果を格納し、前記デコーダーは、前記第１ロー及び第２ロー上で前記ロジック機能を生成し、前記第３ロー内に前記ロジック機能の結果を格納するために、前記少なくとも３つのローのＤＲＡＭ基盤のコンピューティングセルを選択するための命令に対応するＤＲＡＭ基盤のアドレス信号を受信する。

【発明の効果】

【０００９】

本発明によれば、多様な動作に対してプログラム可能且つ再構成可能なＤＰＵを提供することができる。また高性能、エネルギー効率的、低価格のシステムを提供することができる。

【図面の簡単な説明】

【００１０】

【図１】本発明の一実施形態によるＤＰＵの一例を示すブロック図である。

【図２Ａ】コンピューティングセルアレイ内のコンピューティングセルに対して使用される３つのトランジスタ及び１つのキャパシターによるＤＲＡＭコンピューティングセルトポグラフィの一例を示す図である。

【図２Ｂ】コンピューティングセルアレイ内のコンピューティングセルに対して使用される１つのトランジスタ及び１つのキャパシターによるＤＲＡＭコンピューティングセル）トポグラフィの他の例を示す図である。

【図３】本発明の一実施形態によるイントラマットシフトアレイの一例を示す図である。

【図４Ａ】本発明の一実施形態によるインターマットシフトアレイの一例を示す図である。

【図４Ｂ】本発明の一実施形態による左側のインターマットシフトに対する隣接するコンピューティングセルカラムで同様に配置された２つのコンピューティングセル間のインターマットシフトインターコネクション構成を概念的に示す図である。
【図４Ｃ】本発明の一実施形態による左側のインターマットシフトに対する隣接するコンピューティングセルカラムで同一でないように配置された２つのコンピューティングセル間のインターマットシフトインターコネクション構成を概念的に示す図である。
【図５】本発明の一実施形態によるインターマットフォワーディングアレイを示す図である。
【図６Ａ】本発明の一実施形態によるＤＰＵによって提供されるＮＯＲロジック基盤の動作を示す図である。
【図６Ｂ】本発明の一実施形態によるＤＰＵによって提供されるＮＯＲロジック基盤の動作を示す図である。
【図６Ｃ】本発明の一実施形態によるＤＰＵによって提供されるＮＯＲロジック基盤の動作を示す図である。
【図６Ｄ】本発明の一実施形態によるＤＰＵによって提供されるＮＯＲロジック基盤の動作を示す図である。
【図６Ｅ】本発明の一実施形態によるＤＰＵによって提供されるＮＯＲロジック基盤の動作を示す図である。
【図６Ｆ】本発明の一実施形態によるＤＰＵによって提供されるＮＯＲロジック基盤の動作を示す図である。
【図６Ｇ】本発明の一実施形態によるＤＰＵによって提供されるＮＯＲロジック基盤の動作を示す図である。
【図７】本発明の一実施形態による確率的データアレイを含むＤＰＵの一例を示すブロック図である。
【図８Ａ】多重化動作に変換される加算動作に対する確率的コンピューティング動作を示す図である。
【図８Ｂ】ＡＮＤロジック動作に変換される乗算動作に対する確率的コンピューティング動作を示す図である。
【図９】本発明の一実施形態によるＤＰＵを含むシステム構造を示す図である。
【発明を実施するための形態】
【００１１】
以下、本発明を実施するための形態の具体例を、図面を参照しながら詳細に説明する。
【００１２】
本明細書で、多様な詳細な説明を本文の理解を提供するために提示する。しかし、このような詳細な説明無しに記載した本発明の思想が当業者によって容易に具現されることはよく理解される。他の例で、広く公知された方法、手続、構成、及び回路は本文を曖昧にしないために説明しない。
【００１３】
詳細な説明で“１つの実施形態”又は“一実施形態”を参照することは、実施形態に関連する特定の特徴、構造、又は特性が本文の少なくとも１つの実施形態に含まれることを意味する。即ち、本文の多様な箇所で使用する“１つの実施形態で”又は“一実施形態で”又は“１つの実施形態によって”又は類似な意味を有する他の表現は同一の実施形態を参照することを要求しない。更に、特定の特徴、構造、又は特性は適切な方式により１つ以上の実施形態で組合される。これに関連して、本明細書で使用するように、“例示的な”という単語は“例（ｅｘａｍｐｌｅ、ｉｎｓｔａｎｃｅ、ｉｌｌｕｓｔｒａｔｉｏｎ）として提供される”ということを意味する。本明細書で“例示的な”のように記述した実施形態は、他の実施形態に比べて必ずしも望ましいか又は有利であるものと考えてはならない。また、本文で言及する内容に従って、単数用語は複数の形態も含み、複数用語は単数形態も含む。構成図面を含む多様な図面は、説明の目的のためにのみ本文で言及し、正確な縮尺ではない。同様に、多様な波形及びタイミング図は説明の目的のためにのみ図示

する。例えば、一部の構成要素は明確性のために他の構成要素と比較して誇張して図示する。更に、適切に考えられる場合、参照符号は、対応する構成要素及び／又は類似の構成要素を示すために図面で反復する。

【００１４】

本明細書で使用する用語は、特定の実施形態のみを説明するものであり、本発明を制限しようとするものとして意図しない。本明細書で使用するように、文脈上で明確に異なって意味しない限り、単数形態の“１つ”は複数の形態も含むものとして意図する。“構成される”、“構成されている”、“含む”、及び“含んでいる”の用語を本明細書で使用する場合、このような用語は、定まった特徴、整数、段階、動作、要素、及び／又は成分の存在を明示するが、１つ又はそれ以上の他の特徴、整数、段階、動作、要素、成分、及び／又はそれらのグループの追加又は存在を排除しない。“第１”、“第２”のような用語は、先に説明した構成に対するラベルとして使用され、別の定義が無い限り特定の順序（例えば、空間的、時間的、論理的、等）を意図するものではない。更に、同一の参照符号は、同一であるか若しくは類似の機能を有する部品、構成、ブロック、回路、ユニット、又はモジュールに関連する２つ以上の図面に亘って使用される。しかし、このような使用は、説明の簡易化のみのためであり、このような構成及びユニットの構成又は構造的な細部事項が全ての実施形態又は共通的に参照される部分／モジュールで同一なものとして意図せず、本発明の特定の実施形態のみを指称するための１つの手段である。

【００１５】

異なって定義しない限り、全ての用語（技術又は科学用語を含む）は本発明の装置と方法が属する分野で当業者に通常的に理解される同一な意味を有する。また、辞書的な意味として定義した用語は、関連する記述及び／又は本発明の説明の状況に従って解釈されなければならず、このように定義されない限り、理想的に理解されるか若しくは過度に形式的な意味として理解されてはならない。

【００１６】

本明細書に記述する発明は、多様な動作に対してプログラム可能（ｐｒｏｇｒａｍｍａｂｌｅ）であり、再構成可能な（ｒｅｃｏｎｆｉｇｕｒａｂｌｅ）ＤＰＵ（ＤＲＡＭ（Ｄｙｎａｍｉｃ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ） ｂａｓｅｄ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）を提供する。例えば、多様な動作は、加算、乗算、シフト、最大／最小（ＭＩＮ／ＭＡＸ）、及び比較（ＣＯＭＰＡＲＩＳＯＮ）等を含むが、本発明はこれに制限されない。一実施形態で、ＤＰＵは、３つのトランジスタ及び１つのキャパシター（３Ｔ１Ｃ）を含むＤＲＡＭプロセス及び構造に基づく。他の実施形態で、若干の変化と共に、ＤＰＵは、１つのトランジスタ及び１つのキャパシター（１Ｔ１Ｃ）を含むＤＲＡＭプロセス及び構造に基づく。従って、ＤＰＵは、特定のコンピューティングロジック回路（例えば、アダー（加算器）のような）を含まないが、高度の並列動作を使用するメモリセルを使用する計算を提供する。一実施形態で、ＤＰＵは、乗算（ｍｕｌｔｉｐｌｉｃａｔｉｏｎ）動作がＡＮＤロジック動作に変換され、加算（ａｄｄｉｔｉｏｎ）動作が多重化（ｍｕｌｔｉｐｌｅｘｉｎｇ）動作に変換される確率的（ｓｔｏｃｈａｓｔｉｃ）コンピューティングアレイを含む。

【００１７】

また、本明細書に記述する発明は、ＤＰＵをプログラムし、再構成するためのＩＳＡ（Ｉｎｓｔｒｕｃｔｉｏｎ Ｓｅｔ Ａｒｃｈｉｔｅｃｔｕｒｅ）、コンパイラ、ドライバー、ライブラリ、フレームワーク拡張を有する環境（ｅｃｏｓｙｓｔｅｍ）を含むシステム構造を提供する。

【００１８】

更に、本明細書に記述する発明は、データセンター及び／又はモバイルアプリケーションに適合するシステム構造を提供する。そして、システム構造は、ＧＰＵ／ＡＳＩＣ（Ａｐｐｌｉｃａｔｉｏｎ Ｓｐｅｃｉｆｉｃ Ｉｎｔｅｇｒａｔｅｄ Ｃｉｒｃｕｉｔ）（ＴＰＵ）／ＦＰＧＡ（Ｆｉｅｌｄ－Ｐｒｏｇｒａｍｍａｂｌｅ Ｇａｔｅ Ａｒｒａｙ）マシンラーニングアプリケーションに対して選択的に二進（ｂｉｎａｒｙ）及び固定小数

点方式（ｆｉｘｅｄ　ｐｏｉｎｔ）の計算の両方に対するマシンラーニングアプリケーションのためのＰＩＭ（Ｐｒｏｃｅｓｓｏｒ－ｉｎ　Ｍｅｍｏｒｙ）ソリューションを提供する。一実施形態で、本明細書に記述する発明は、高性能、エネルギー効率的、低価格のシステムを提供する。該当システムは、例えば二進加重ニューラルネットワーク（Ｂｉｎａｒｙ　Ｗｅｉｇｈｔ　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）に対する加速化されたディープラーニングを提供する。

【００１９】

本明細書に記述する発明は、再構成及びプログラム可能であり、ＤＲＡＭ技術を利用して構成されるＤＰＵに関連する。一実施形態で、ＤＰＵは、多様な動作（例えば、加算、乗算、整列、等）を遂行するように構成されるＤＲＡＭ基盤のメモリセルアレイ及びＤＲＡＭ基盤のコンピューティングセルアレイを含む。

【００２０】

ＤＰＵの内部構成は、サブアレイの複数のバンクに連結されたシステムバスを含む。一実施形態で、システムバスはサブアレイのＨツリー連結バンクを提供するように構成される。各々のサブアレイはローカルコントローラを含み、各々の独立的なサブアレイは分離的に又は同時に活性化される。一実施形態で、ＤＲＡＭ基盤のセルは２つのアレイ（データセルアレイ及びコンピューティングセルアレイ）に区分される。一実施形態で、コンピューティングセルアレイはＤＲＡＭ基盤のメモリセルで具現される。他の実施形態で、コンピューティングセルアレイはロジック回路を含むＤＲＡＭ基盤のメモリセルで具現される。ＤＰＵ内部構造は、またデータシフト及びデータ移動回路を含む。一実施形態で、確率的データ計算のために構成される第３ＤＲＡＭ基盤のセルアレイがある。

【００２１】

図１は、本発明の一実施形態によるＤＰＵ（ＤＲＡＭ（Ｄｙｎａｍｉｃ　Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）　ｂａｓｅｄ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）１００の一例を示すブロック図である。ＤＰＵ１００は１つ以上のバンク（１０１ａ～１０１ｍ）を含む。図１では、１つ以上のバンク（１０１ａ～１０１ｍ）の中のバンク１０１ａ、１０１ｂのみを図示する。各バンク１０１は１つ以上のサブアレイ（１０２ａ～１０２ｎ）を含む。図１では、１つ以上のサブアレイ（１０２ａ～１０２ｎ）の中のサブアレイ１０２ａ、１０２ｂのみを図示する。また、各バンク１０１はバッファ１０３を含む。バッファ１０３は、個別サブアレイ１０２に連結され、システムバス１０４に連結される。バッファ１０３は、バンク１０２内の全体ロー（ｒｏｗ）を読み出し、その後該当ローを再び同一のバンク又は他のバンクに書き込む。バッファ１０３は、また該当ローデータのコピーをサブアレイ１０２内の複数のマット（１０５ａ～１０５ｎ）にブロードキャスト（ｂｒｏａｄｃａｓｔ）する。一実施形態で、バンク１０１及びシステムバス１０４はＨツリー連結バンクを提供する。

【００２２】

各サブアレイ１０２は１つ以上のマット（又はレーン）１０５を含む。図１では、１つ以上のマット１０５の中のサブアレイ１０２ａのマット１０５ａ～１０５ｎのみを図示する。各マット１０５は、データセルアレイ１０６、コンピューティングセルアレイ１０７、及びイントラマット（ｉｎｔｒａ－ｍａｔ）シフトアレイ１０８を含むＤＰＵ１００の一領域である。マット１０５の例は点線１０９で囲まれるように図１で示される。各マット１０５は、データセルアレイデコーダー１１０、コンピューティングセルアレイデコーダー１１１、インターマット（ｉｎｔｅｒ－ｍａｔ）シフトアレイ１１２、及びインターマットフォワーディングアレイ１１３を、隣接するマットと共有する。一実施形態で、データセルアレイデコーダー１１０、コンピューティングセルアレイデコーダー１１１、及びインターマットシフトアレイ１１２は、隣接するマット１０５の間にサブアレイコントローラ１１４と交互に物理的に配置される。一実施形態で、各デコーダー（１１０、１１１）は従来のＤＡＲＭタイプのメモリデコーダーのように動作する。

【００２３】

一実施形態で、各マット１０５は通信的に（ｃｏｍｍｕｎｉｃａｔｉｖｅｌｙ）サブア

レイコントローラ１１４に連結される。各サブアレイコントローラ１１４は他のサブアレイコントローラ１１４から独立して構成される。サブアレイコントローラ１１４はアドレス（ａｄｄｒ）として命令をＤＲＡＭアドレスバスから受信する。アドレス（即ち、アドレス信号）に応答して、サブアレイコントローラ１１４は、データセルアレイ１０６及びコンピューティングセルアレイ１０７の中の少なくとも１つのアレイに出力としてデコーディングされたアドレスを提供する。即ち、サブアレイコントローラ１１４は、関連するデータセルアレイ１０６に対するデータセルアレイデコーダー１１０によってデコーディングされたソース／目的地（ｄｅｓｔｉｎａｔｉｏｎ）（ｓｒｃ／ｄｓｔ）アドレスを出力する。そして、コンピューティングセルアレイ１０７の場合、サブアレイコントローラ１１４は、コンピューティングセルアレイデコーダー１１１によってデコーディングされた動作／計算（ｏｐ／ｃａｌｃ）アドレスを出力する。また、サブアレイコントローラ１１４は、２つ以上のサブアレイコントローラ１１４が組織化された方法で動作するＤＲＡＭバスからのアドレスとして命令を受信する。また、サブアレイコントローラ１１４はデータ移動回路を制御する。例えば、サブアレイコントローラ１１４は、イントラマットシフトアレイ１０８、インターマットシフトアレイ１１２、及びインターマットフォワーディングアレイ１１３を制御する。

【００２４】

各データセルアレイ１０６は、少なくとも１つのカラム（ｃｏｌｕｍｎ）及び少なくとも１つのロー（ｒｏｗ）内に配置される１つ以上のＤＲＡＭセルを含む。一実施形態で、データセルアレイ１０６は従来のＤＲＡＭセルアレイのように構成される。一実施形態で、データセルアレイ１０６は２Ｋカラム及び１６ローを含む。他の実施形態で、データセルアレイ１０６は２Ｋカラムより少ないか又は多いカラムを含み、また１６ローより少ないか又は多いローを含む。

【００２５】

各コンピューティングセルアレイ１０７は、少なくとも１つのカラム及び少なくとも１つのロー内に配置される１つ以上のコンピューティングセルを含む。コンピューティングセルアレイ１０７内のカラム数はデータセルアレイ１０６内のカラム数と同一である。一実施形態で、コンピューティングセルアレイ１０７は２Ｋカラム及び１６ローを含む。他の実施形態で、コンピューティングセルアレイ１０７は２Ｋカラムより少ないか又は多いカラムを含み、また１６ローより少ないか又は多いローを含む。

【００２６】

図２Ａは、コンピューティングセルアレイ１０７内のコンピューティングセルに対して使用される３つのトランジスタ及び１つのキャパシター（３Ｔ１Ｃ）によるＤＲＡＭコンピューティングセルトポグラフィ（ｔｏｐｏｇｒａｐｈｙ）２０１の一例を示す図である。図２Ａに示すように、ローＸ内の３Ｔ１Ｃコンピューティングセルは第１トランジスタＴ１を含む。第１トランジスタＴ１は、書込みビットライン（Ｗｒｉｔｅ　ＢＬ）に電気的に連結されたソース端、キャパシターＣ１の第１端及び第２トランジスタＴ２のゲート端の両側に電気的に連結されたドレーン端、並びに書込みイネーブル（ＷＥＮ）ラインに電気的に連結されたゲート端を含む。キャパシターＣ１の第２端はグラウンド（ｇｒｏｕｎｄ）ラインに電気的に連結される。第２トランジスタＴ２は、グラウンドラインに電気的に連結されたソース端、及び第３トランジスタＴ３のソース端に電気的に連結されたドレーン端を含む。第３トランジスタＴ３は、ワードラインＷＬに電気的に連結されたゲート端、及び読出しビットライン（Ｒｅａｄ　ＢＬ）に電気的に連結されたドレーン端を含む。３Ｔ１Ｃコンピューティングセルトポグラフィ２０１は、読出しビットライン（Ｒｅａｄ　ＢＬ）に電気的に連結された入力及び書込みビットライン（Ｗｒｉｔｅ　ＢＬ）に電気的に連結された出力を含むセンスアンプ（Ｓｅｎｓｅ　Ａｍｐｌｉｆｉｅｒ：ＳＡ）を含む。

【００２７】

また、ローＹ内のコンピューティングセル及びローＲ内のコンピューティングセルの両方は、ローＸ内のコンピューティングセルの配置と同様に、３Ｔ１Ｃ　ＤＲＡＭ構成に配

置された３つのトランジスタ（Ｔ１～Ｔ３）及びキャパシターＣを含む。図２Ａに示した３つのコンピューティングセル及びセンスアンプ（ＳＡ）の一例は、ＮＯＲロジック動作（即ち、‘Ｘ　ＮＯＲ　Ｙ’ロジック動作）を提供するように構成される。該当ＮＯＲロジック動作の結果はローＲに格納される。３Ｔ１Ｃ　ＤＲＡＭコンピューティングセルの１つのカラムのみを図２Ａに明示的に示したが、３Ｔ１Ｃコンピューティングセルが複数のカラム（例えば、２Ｋカラム）内に構成される等の他の実施形態が可能である。また、３つ以上のローが提供される他の実施形態が可能である。また、図２Ａに示した３Ｔ１Ｃ　ＤＲＡＭコンピューティングセル構成はＮＯＲロジック動作を提供するが、３Ｔ１Ｃ　ＤＲＡＭコンピューティングセルトポグラフィ２０１のＮＯＲロジック動作は多様な機能的な動作を提供するために使用される。例えば、機能的な動作は、排他ノア（ＸＮＯＲ）、加算（ＡＤＤ）、セレクト（ＳＥＴ）、ＭＡＸ、ＳＩＧＮ、多重化（ＭＵＸ）、ＣＳＡ（Ｃｏｎｄｉｔｉｏｎａｌ　Ｓｕｍ　Ａｄｄｉｔｉｏｎ）ロジック、乗算、ポップカウント（ｐｏｐｃｏｕｎｔ）、ＣＯＭＰＡＲＥ等を含む。但し、本発明はこれに制限されない。また、イントラマットシフトアレイ１０８及びインターマットシフトアレイ１１２はシフト機能を提供する。

【００２８】

図２Ｂは、図１のコンピューティングセルアレイ１０７内のコンピューティングセルに対して使用される１つのトランジスタ及び１つのキャパシター（１Ｔ１Ｃ）によるＤＲＡＭコンピューティングセルトポグラフィ（ｔｏｐｏｇｒａｐｈｙ）２０２の他の例を示す図である。図２Ｂに示すように、１Ｔ１Ｃコンピューティングセルは第４トランジスタＴ４を含む。第４トランジスタＴ４は、キャパシターＣ２の第１端に電気的に連結されたソース端、ビットラインＢＬに電気的に連結されたドレーン端、及びワードラインＷＬに電気的に連結されたゲート端を含む。キャパシターＣ２の第２端はグラウンドラインに電気的に連結される。ビットラインＢＬはセンスアンプ（ＳＡ）の入力に電気的に連結される。センスアンプ（ＳＡ）の出力は、多重化器（ＭＵＸ）の第１入力、第５トランジスタＴ５のドレーン端、及びＡＬＵ（Ａｒｉｔｈｍｅｔｉｃ　Ｌｏｇｉｃ　Ｕｎｉｔ）の入力に電気的に連結される。多重化器（ＭＵＸ）の出力はラッチ（ＬＡＴＣＨ）の入力に電気的に連結される。第５トランジスタＴ５のソース端はラッチ（ＬＡＴＣＨ）の出力に電気的に連結される。ＡＬＵの出力は多重化器（ＭＵＸ）の第２入力に電気的に連結される。図２Ｂで、第５トランジスタＴ５、多重化器（ＭＵＸ）、ラッチ（ＬＡＴＣＨ）、及びＡＬＵは、それぞれコントローラ１１４から制御信号（ＣＮＴＬ１～ＮＴＬ４）を受信する。一実施形態で、ＡＬＵはＮＯＲ機能を提供するように構成される。図２Ｂのビットライン ＢＬに電気的に連結されたロジック回路はＮＯＲロジック動作を提供するが、ビットライン ＢＬに電気的に連結されたロジック回路（即ち、ＡＬＵ）は他の機能動作（例えば、排他ノア（ＸＮＯＲ）、加算（ＡＤＤ）、セレクト（ＳＥＴ）、ＭＡＸ、ＳＩＧＮ、多重化（ＭＵＸ）、ＣＳＡ（Ｃｏｎｄｉｔｉｏｎａｌ　Ｓｕｍ　Ａｄｄｉｔｉｏｎ）ロジック、乗算、ポップカウント（ｐｏｐｃｏｕｎｔ）、ＣＯＭＰＡＲＥ等）を提供する。但し、本発明はこれに制限されない。また、イントラマットシフトアレイ１０８及びインターマットシフトアレイ１１２はシフト機能を提供する。図２Ｂで１つの１Ｔ１Ｃコンピューティングセルのみを図示したが、複数のカラム及びローの１Ｔ１Ｃコンピューティングセルが提供される。

【００２９】

図２Ａ及び図２Ｂから分かるように、ＤＰＵのコンピューティングセルは、特定の複雑なコンピューティングロジックを含まない。但し、代わりに、ＤＰＵのコンピューティングセルは、複数の多様なタイプの計算を遂行する機能（ａｂｉｌｉｔｙ）を提供する再プログラム可能な性質（ｎａｔｕｒｅ）を有する相対的に単純なトポグラフィを含む。また、ＤＰＵのポグラフィは、より多くの計算をより速くより効率的に遂行するためにメモリ構造に内在され、大量並列処理の長所を有するように配置される。

【００３０】

図３は、本発明の一実施形態によるイントラマット（ｉｎｔｒａ－ｍａｔ）シフト（ｓ

ｈｉｆｔ）アレイ１０８の一例を示す図である。イントラマットシフトアレイ１０８の記述（ｄｅｓｃｒｉｐｔｉｏｎ）を単純化するために、図３に示すように、４つのカラムのコンピューティングメモリセルアレイ１０７の幅のマット１０５を考慮する。イントラマットシフトアレイ１０８は、アレイ内に配置された複数の第６トランジスタＴ６（図３では、１つのトランジスタのみをＴ６で表示する）、2^nシフトラインＳＬｓ（ｎはマット１０５内のコンピューティングセルのカラムである）、ｎ＋２シフトレフト（ｌｅｆｔ）コントロールラインＳＬｃＬｓ、２シフトライト（ｒｉｇｈｔ）コントロールラインＳＲｃＬｓ、及びｎシフトマスクラインＳＭＬｓを含む。イントラマットシフトアレイ１０８の第６トランジスタＴ６の一部は書込みビットライン（Ｗｒｉｔｅ　ＢＬ）と2^nシフトラインＳＬｓとの間に電気的に連結され、イントラマットシフトアレイ１０８の他の第６トランジスタＴ６は読出しビットライン（Ｒｅａｄ　ＢＬ）と2^nシフトラインＳＬｓとの間に連結される。このような第６トランジスタＴ６のゲートはｎ＋２シフトレフトコントロールラインＳＬｃＬｓ及び２シフトライトコントロールラインＳＲｃＬｓに電気的に連結される。イントラマットシフトアレイ１０８の他の第６トランジスタＴ６はｎシフトマスクラインＳＭＬｓと2^nシフトラインＳＬｓとの間に電気的に連結される。イントラマットシフトアレイ１０８のコントロールラインはマット１０５に関連するサブアレイコントローラ１１４に電気的に連結される。

【００３１】

コントロールライン（ＳＬｃＬｓ、ＳＲｃＬｓ）上の適切な信号によって、イントラマットシフトアレイ１０８は、マット１０５内でデータをレフト（ｌｅｆｔ）シフト又はライト（ｒｉｇｈｔ）シフトする。レフトシフトに対して、データは、サイン（ｓｉｇｎ）ビットで満たされ、１つの動作毎に１ビット又は（ｎ－１）ビット程シフトされる。ｎはマット１０５当たりのカラム数である。ライトシフトに対して、命令による制御に従ってデータは０又は１で満たされる。或いは、データは2^0、2^1、…、2^{k-1}、2^kマット当たりのカラム数までシフトされる。2^kはカラム数である。

【００３２】

図４Ａは、本発明の一実施形態によるインターマット（ｉｎｔｅｒ－ｍａｔ）シフトアレイ１１２の一例を示す図である。インターマットシフトアレイ１１２の記述（ｄｅｓｃｒｉｐｉｔｉｏｎ）を単純化するために、図４Ａ～図４Ｃに示すように、マット１０５が２つのカラムのコンピューティングメモリセルアレイ１０７の幅であるマット１０５の構成を考慮する。即ち、各マット１０５はコンピューティングメモリセル１０７ａの第１カラム及びコンピューティングメモリセル１０７ｂの第２カラムを含む。インターマットシフトアレイ１１２は、トランジスタＴ１１２ａ、Ｔ１１２ｂ、トランジスタＴ１１２ｃ、Ｔ１１２ｄ、データシフトライン１１２ｅ、１１２ｆ、及びインターマットシフトコントロールラインＩＳＬｃＬｓを含む。マット内で、トランジスタＴ１１２ａは、コンピューティングメモリセル１０７ａの第１カラムの読出しビットライン（Ｒｅａｄ　ＢＬ）に電気的に連結されたソース端、データシフトライン１１２ｅに電気的に連結されたドレーン端を含む。トランジスタＴ１１２ｂは、コンピューティングメモリセル１０７ｂの第２カラムの読出しビットライン（Ｒｅａｄ　ＢＬ）に電気的に連結されたソース端、データシフトライン１１２ｆに電気的に連結されたドレーン端を含む。データシフトライン（１１２ｅ、１１２ｆ）はバッファ１０３に電気的に連結される。バッファ１０３は図４Ａに図示していない。異なるマットの間で、トランジスタＴ１１２ｃは、隣接マット内のデータシフトライン１１２ｅにそれぞれ電気的に連結されたソース端及びドレーン端を含む。トランジスタＴ１１２ｄは、隣接マット内のデータシフトライン１１２ｆにそれぞれ電気的に連結されたソース端及びドレーン端を含む。トランジスタＴ１１２ｃ、Ｔ１１２ｄのゲートはそれぞれ異なるインターマットシフトコントロールラインＩＳＬｃＬｓのそれぞれに連結される。インターマットシフトコントロールラインＩＳＬｃＬｓ上の適切な信号によって、インターマットシフトアレイ１１２は、異なるマットの間でデータをレフトシフト又はライトシフトする。インターマットシフトアレイ１１２のコントロールラインはマット１０５に関連するサブアレイコントローラ１１４に電気的に連結される。

【００３３】
図４Ｂは、本発明の一実施形態による左側のインターマットシフトに対する隣接するコンピューティングセルカラム（マット１０５ａ、１０５ｂ）で同様に配置された２つのコンピューティングセル間のインターマットシフトインターコネクション（ｉｎｔｅｒｃｏｎｎｅｃｔｉｏｎ）構成を概念的に示す図である。図４Ｂのインターコネクション構成を、利用されるインターコネクションノードによって概念的に図示し、利用されるインターコネクションノードを強調して図示する。例えば、トランジスタＴ１１２ｃ、Ｔ１１２ｄは活性化され、これに従って導線経路が各々のトランジスタの間に形成される。従って、左側のコンピューティングセルカラム（マット１０５ａ）と右側のコンピューティングセルカラム（マット１０５ｂ）との間でデータシフトライン（１１２ｅ、１１２ｆ）は連結される。トランジスタＴ１１２ｃ、Ｔ１１２ｄのゲート端はアクティブなインターマットシフトコントロールラインＩＳＬｃＬに電気的に連結される。マット１０５ｂ内のトランジスタＴ１１２ａ、Ｔ１１２ｂは活性化され、従ってマット１０５ｂ内のコンピューティングセル１０７ａの読出しビットライン（Ｒｅａｄ　ＢＬ）は、マット１０５ｂの左側であるマット１０５ａ内のコンピューティングセル１０７ａの書込みビットライン（Ｗｒｉｔｅ　ＢＬ）に電気的に連結される。また、これに従って、マット１０５ｂ内のコンピューティングセル１０７ｂの読出しビットライン（Ｒｅａｄ　ＢＬ）は、マット１０５ｂの左側であるマット１０５ａ内のコンピューティングセル１０７ｂの書込みビットライン（（Ｗｒｉｔｅ　ＢＬ）に電気的に連結される。

【００３４】
図４Ｃは、本発明の一実施形態による左側のインターマットシフトに対する隣接するコンピューティングセルカラム（１０５ａ、１０５ｂ）で同一でないように配置された２つのコンピューティングセル間のインターマットシフトインターコネクション（ｉｎｔｅｒｃｏｎｎｅｃｔｉｏｎ）構成を概念的に示す図である。図４Ｃのインターコネクション構成を、利用されるインターコネクションノードによって概念的に図示し、利用されるインターコネクションノードを強調して図示する。例えば、トランジスタＴ１１２ｃ、Ｔ１１２ｄは活性化されて導線経路が各々のトランジスタの間に形成される。従って、左側のコンピューティングセルカラム（マット１０５ａ）と右側のコンピューティングセルカラム（マット１０５ｂ）との間でデータシフトライン（１１２ｅ、１１２ｆ）は連結される。トランジスタＴ１１２ｃ、Ｔ１１２ｄのゲート端はアクティブなインターマットシフトコントロールラインＩＳＬｃＬに電気的に連結される。マット１０５ａのトランジスタＴ１１２ａ、Ｔ１１２ｂは活性化され、従ってマット１０５ａ内のコンピューティングセル１０７ａの読出しビットライン（Ｒｅａｄ　ＢＬ）は、マット１０５ａの右側であるマット１０５ｂ内のコンピューティングセル１０７ａの書込みビットライン（Ｗｒｉｔｅ　ＢＬ）に電気的に連結される。また、これに従って、マット１０５ａ内のコンピューティングセル１０７ｂの読出しビットライン（Ｒｅａｄ　ＢＬ）は、マット１０５ａの右側であるマット１０５ｂ内のコンピューティングセル１０７ｂの書込みビットライン（Ｗｒｉｔｅ　ＢＬ）に電気的に連結される。

【００３５】
図５は、本発明の一実施形態によるインターマットフォワーディング（ｆｏｒｗａｄｉｎｇ）アレイ１１３を示す図である。インターマットフォワーディングアレイ１１３の説明を単純化するために、図５に示すように、２つのカラムのコンピューティングセルアレイ１０７の幅のマット１０５の構成を考慮する。即ち、各マット１０５はコンピューティングセル１０７ａの第１カラム及びコンピューティングセル１０７ｂの第２カラムを含む。マット１０５と共に、インターマットフォワーディングアレイ１１３は、トランジスタＴ１１３ａ、Ｔ１１３ｂ、トランジスタＴ１１３ｃ、Ｔ１１３ｄ、トランジスタＴ１１３ｅ、Ｔ１１３ｆ、2^{n}フォワーディングデータラインＦＤＬ、フォワーディングコントロールラインＦＣＬ、及び2^{m}フォワーディングセクションラインＦＳＬを含む。ここで、ｎはマット内のコンピューティングセルのカラム数であり、ｍはセクション数である。トランジスタＴ１１３ａ、Ｔ１１３ｂのソース端はそれぞれコンピューティングセル１０７

ａの第１カラムの書込みビットライン（Ｗｒｉｔｅ　ＢＬ）及び読出しビットライン（Ｒｅａｄ　ＢＬ）に電気的に連結される。トランジスタＴ１１３ａ、Ｔ１１３ｂのドレーン端は第１データフォワーディングライン（ＦＤＬ）１１３ｇに電気的に連結される。トランジスタＴ１１３ｃ、Ｔ１１３ｄのソース端はそれぞれコンピューティングセル１０７ｂの第２カラムの書込みビットライン（Ｗｒｉｔｅ　ＢＬ）及び読出しビットライン（Ｒｅａｄ　ＢＬ）に電気的に連結される。トランジスタＴ１１３ｃ、Ｔ１１３ｄのドレーン端は第２データフォワーディングライン（ＦＤＬ）１１３ｈに電気的に連結される。トランジスタＴ１１３ｅ、Ｔ１１３ｆのソース端はそれぞれトランジスタＴ１１３ａ、Ｔ１１３ｂのゲート端に電気的に連結される。トランジスタＴ１１３ｅ、Ｔ１１３ｆのドレーン端は両方とも同一のフォワーディングセクションラインＦＳＬに連結される。トランジスタＴ１１３ｅ、Ｔ１１３ｆのゲート端はそれぞれ異なるフォワーディングコントロールラインＦＣＬｓに連結される。フォワーディングコントロールラインＦＣＬｓ上の適切な信号によって、インターマットフォワーディングアレイ１１３は、マットの間でデータをフォワーディングする。インターマットフォワーディングアレイ１１３のコントロールラインは、相互間でデータがフォワーディングされるマット１０５に関連するサブアレイコントローラ１１４に電気的に連結される。

【００３６】

図６Ａ～図６Ｇは、本発明の一実施形態によるＤＰＵによって提供されるＮＯＲロジック基盤の動作を示す図である。図６Ａ～図６Ｇで、第１オペランド（ｏｐｅｒａｎｄ）はローＸに格納され、第２オペランドはローＹ又はローＷに格納される。図６Ａ～図６Ｇ内の矢印はコンピューティングセルの全体ローに対するＮＯＲロジック動作の入出力フローを示す。例えば、図６Ａのローはローのコンピューティングセルに格納されたオペランドの全体ローを示す。ローＸ内に格納されたオペランド及びローＹ内に格納されたオペランドのＮＯＲロジック動作の結果は結果ローＲ内に格納される。一実施形態で、ローＸ及びローＹのオペランドは、例えば１００個のカラム（例えば、x_1、x_2、…、x_{100}及びy_1、y_2、…、y_{100}）を含み、結果はローＲ（例えば、r_1、r_2、…、r_{100}）内に格納される。即ち、‘x_i　ＮＯＲ　$y_i = r_i$’である。ここで、ｉはカラムインデックスである。他の実施形態で、ローＸはロー内のコンピューティングセルの選択されたグループのみを示す。

【００３７】

図６Ｂはプレフィックス（ｐｒｅｆｉｘ）Ｋｏｇｇｅ－Ｓｔｏｎｅアダー（加算器）に基づくＮビット数に対するプールアダー動作を例示的に示す。図６Ｂで、第１ＮビットオペランドはローＸ内に格納され、第２ＮビットオペランドはローＹ内に格納される。図６Ｂ内に示した例示的な加算動作で、中間値（ｉｎｔｅｒｍｅｄｉａｔｅ　ｔｅｒｍ）（G_0、P_0、G_1、P_1、G_2、P_2、…、G_{logN+1}、P_{logN+1}）が計算される。図６Ｂの最上のブロックはローＸ及びローＹからの入力オペランドを利用してG_0及びP_0を決定する５個の分離された動作を示す。第１動作で、最上のブロックはローＸの逆（即ち、～Ｘ）を決定し、これはロー１に格納される。第２動作はローＹの逆（即ち、～Ｙ）を決定し、これはロー２に格納される。第３動作は‘ローＸ　ＮＯＲ　ローＹ’の動作を決定し、これはロー３に格納される。第４動作は‘Ｇ０＝ロー１　ＮＯＲ　ロー２’を決定し、これはロー４に格納される。第５動作は‘P_0＝ロー３　ＮＯＲ　ロー４’を決定し、これはロー５に格納される。

【００３８】

図６Ｂの中間ブロックを参照すると、最上ブロックからの中間結果G_0、P_0が中間結果G_{i+1}、P_{i+1}を決定するために使用される。ここで、ｉはカラムインデックスである。即ち、図６Ａの最上のブロックから決定された中間結果G_0、P_0が中間結果G_1、P_1を決定するために使用される。中間結果G_1、P_1は中間結果G_2、P_2を決定するために使用され、同様に中間結果G_{logN+1}、P_{logN+1}が決定される。図６Ｂの最下のブロックで、結果ローＲ１及び結果ローＲ２はそれぞれプールアダー（ｆｕｌｌ　ａｄｄｅｒ）動作に対するキャリー（ｃａｒｒｙ）結果及び総合（ｓｕｍ）結果を格

納する。

【００３９】

図６Ｃは３Ｔ１Ｃ　ＤＲＡＭコンピューティングセルトポグラフィ２０１によって提供される例示的なセレクター動作を示す。ロー１はローＸの逆（即ち、～Ｘ）の中間結果を格納する。ロー２はローＹの逆（即ち、～Ｙ）の中間結果を格納する。ロー３はローＳの逆（即ち、～Ｓ）の中間結果を格納する。ロー４は‘ロー１　ＮＯＲ　ロー３’の中間結果を格納する。ロー５は‘ロー２　ＮＯＲ　ローＳ’の中間結果を格納する。ロー６は‘ロー４　ＮＯＲ　ロー５’の中間結果を格納する。ローＲはロー６の逆の結果（Ｓ？Ｘ：Ｙ）を格納する。

【００４０】

図６Ｄは３Ｔ１Ｃ　ＤＲＡＭコンピューティングセルトポグラフィ２０１によって提供される他の例示的なセレクター動作を示す。ロー１はローＸの逆（即ち、～Ｘ）の中間結果を格納する。ロー２はローＳの逆（即ち、～Ｓ）の中間結果を格納する。ロー３は‘ロー１　ＮＯＲ　ローＳ’の中間結果を格納する。ロー４はローＸの逆（～Ｘ）の中間結果を格納する。ローＲは‘ロー３　ＮＯＲ　ロー４’の結果（Ｓ？Ｘ：～Ｘ）を格納する。

【００４１】

図６Ｅは３Ｔ１Ｃ　ＤＲＡＭコンピューティングセルトポグラフィ２０１によって提供される例示的なＭＡＸ／ＭＩＮ動作を示す。ロー１はローＹの逆（即ち、～Ｙ）の中間結果を格納する。ロー２はローＸ＋（～Ｙ＋１）の中間結果を格納する。ロー３は‘C_{out}＞＞ｎ’の中間結果を格納する。ロー４は‘C_{out}？Ｘ：Ｙ’の中間結果を格納する。ローＲは‘ＭＡＸ（Ｘ：Ｙ）’の結果を格納する。

【００４２】

図６Ｆは３Ｔ１ＣＤＲＡＭコンピューティングセルトポグラフィ２０１によって提供される例示的な１ビット乗算動作を示す。ロー１は‘ローＸ　ＮＯＲ　ローＷ’の中間結果を格納する。ロー２は‘ローＸ　ＮＯＲ　ロー１’の中間結果を格納する。ロー３は‘ローＷ　ＮＯＲ　ロー１’の中間結果を格納する。結果ローＲは‘ロー２　ＮＯＲ　ロー３’の結果（即ち、‘ローＸ　ＸＮＯＲ　ローＷ’の結果）を格納する。

【００４３】

図６Ｇは３Ｔ１Ｃ　ＤＲＡＭコンピューティングセルトポグラフィ２０１によって提供される例示的な複数ビット乗算動作を示す。図６Ｇの上方のブロックで、ロー１はローＷの逆（即ち、～Ｗ）の中間結果を格納する。ロー２は2^{i}番目のレフトシフトされたローＸの逆（即ち、～Ｘ＜＜2^{i}）の中間結果を格納する。ここで、ｉはインデックスである。ロー３は‘ロー１　ＮＯＲ　ロー２’の中間結果（PP_{i}＝～Ｗ　ＮＯＲ　～Ｘ＜＜2^{i}）を格納する。図６Ｇの下方のブロックで、ロー１は‘ローPP_{0}　ＳＵＭ　ローPP_{i}（ΣPP_{i}）’の中間結果を格納する。ロー２は‘ロー２　ＮＯＲ　ローW_{sign}’の中間結果を格納する。ローＲは‘Ｘ＊Ｗ’の結果を格納する。

【００４４】

図７は、本発明の一実施形態による確率的（ｓｔｏｃｈａｓｔｉｃ）データアレイ７１５を含むＤＰＵ７００の一例を示すブロック図である。図１に示したＤＰＵ１００の構成要素と同一な参照インジケータ（ｉｎｄｉｃａｔｏｒ）を有するＤＰＵ７００の多様な構成要素は同様であり、このような同様の構成要素の記述はここで省略する。ＤＰＵ７００のサブアレイ１０２は、（実際の）データセルアレイ１０６、コンピューティングセルアレイ１０７、及びイントラマットシフトアレイ１０８と共に、確率的データアレイ７１５及びコンバーターツー確率（ｃｏｎｖｅｒｔｅｒ－ｔｏ－ｓｔｏｃｈａｓｔｉｃ）アレイ７１６を含む。

【００４５】

確率的データアレイ７１５の各々は、少なくとも１つのカラム及び少なくとも１つのロー内に配置された１つ以上の確率的コンピューティングセルを含む。確率的データアレイ７１５内のカラム数はデータセルアレイ１０６及びコンピューティングセルアレイ１０７内のカラム数と同一である。一実施形態で、確率的データアレイ７１５は２Ｋカラム及び

１６ローを含む。他の実施形態で、確率的データアレイ７１５は２Ｋカラムより少ないか又は多いカラム、或いは１６ローより少ないか又は多いローを含む。確率的データアレイ７１５内で、‘１’が存在する確率が使用され、２ｎビットはｎビット値を示すために使用される。コンバーターツー確率アレイ７１６内の任意数生成器は実数（ｒｅａｌ　ｎｕｍｂｅｒ）を確率的な数に変換するために使用される。ポップカウント動作は確率的な数を再び実数に変換するために使用される。

【００４６】

確率的なコンピューティングアプローチを使用して、加算は多重化（ｍｕｌｔｉｐｌｅｘｉｎｇ）動作に変換され、乗算はＡＮＤロジック動作に変換される。例えば、図８Ａは、多重化動作に変換される加算動作に対する確率的コンピューティング動作を示す図であり、図８Ｂは、ＡＮＤロジック動作に変換される乗算動作に対する確率的コンピューティング動作を示す図である。確率的コンピューティングに対する従来の技術は巨大なメモリ容量を要求する。しかし、本明細書に記述した発明は高度に効率的な確率的コンピューティングを提供するために使用される。これはＤＲＡＭ基盤のＤＰＵが多くの並列ＡＮＤ動作及びマックス（ＭＵＸ）動作を遂行するためである。本明細書に記述したＤＰＵを使用する確率的コンピューティングは、またディープラーニングが一般的なアプリケーションである複雑な動作を加速化する。

【００４７】

図９は、本発明の一実施形態によるＤＰＵを含むシステム構造９００を示す図である。システム構造９００は、ハードウェアレイヤー９１０、ライブラリ及びドライバーレイヤー９２０、フレームワークレイヤー９３０、及びアプリケーションレイヤー９４０を含む。

【００４８】

ハードウェアレイヤー９１０は、本明細書に記述したＤＰＵのような内装されたＤＰＵを含むハードウェア装置及び／又は構成要素を含む。装置及び／又は構成要素の一実施形態は、１つ以上の内装されたＤＰＵを含むＰＣＩｅ装置９１１である。装置及び／又は構成要素の他の実施形態は、１つ以上の内装されたＤＰＵを含むＤＩＭＭ（Ｄｕａｌ　Ｉｎ－ｌｉｎｅ　Ｍｅｍｏｒｙ　Ｍｏｄｕｌｅ）９１２である。システム構造９００のハードウェアレイヤー９１０はＰＣＩｅ装置及び／又はＤＩＭＭに制限されず、ハードウェアレイヤー９１０はＤＰＵを含むＳＯＣ（Ｓｙｓｔｅｍ　Ｏｎ　Ｃｈｉｐ）装置又は他のメモリタイプ装置を含むことは容易に理解される。ハードウェアレベル９１０で装置及び／又は構成要素内に内装されるＤＰＵは、図１のＤＰＵ１００及び／又は図７のＤＰＵ７００と同様に構成される。他の実施形態で、ＤＰＵの特定コンピューティングセルアレイは、図２Ａの３Ｔ１Ｃコンピューティングセルトポグラフィ２０１又は図２Ｂの１Ｔ１Ｃコンピューティングセルトポグラフィ２０２を含むように構成される。

【００４９】

システム構造９００のライブラリ及びドライバーレイヤー９２０は、ＤＰＵライブラリ９２１、ＤＰＵドライバー９２２、及びＤＰＵコンパイラ９２３を含む。ＤＰＵライブラリ９２１は、アプリケーションレイヤー９４０で動作する多様なアプリケーションに対するハードウェアレイヤー９１０内のＤＰＵ内サブアレイの各々に対する最適のマッピング機能、リソース割当機能、及びスケジューリング機能を提供するように構成される。

【００５０】

一実施形態で、ＤＰＵライブラリ９２１は、移動、加算、乗算等の動作を含むフレームワークレイヤー９３０に対するハイレベルＡＰＩ（Ａｐｐｌｉｃａｔｉｏｎ　Ｐｒｏｇｒａｍｍｉｎｇ　Ｉｎｔｅｒｆａｃｅ）を提供する。例えば、ＤＰＵライブラリ９２１は、また標準タイプのルーチン（ｒｏｕｔｉｎｅｓ）に対する具現を含む。標準タイプのルーチンに対する具現は、加速化されたディープラーニングプロセスに対して適用されるフォワード（ｆｏｒｗａｒｄ）及びバックワード（ｂａｃｋｗａｒｄ）コンボリューション（ｃｏｎｖｏｌｕｔｉｏｎ）、プーリング（ｐｏｏｌｉｎｇ）、正規化（ｎｏｒｍａｌｉｚａｔｉｏｎ）、及び活性化（ａｃｔｉｖａｔｉｏｎ）レイヤーを含む。但し、本発明はこ

れに制限されない。一実施形態で、ＤＰＵライブラリ９２１は、ＣＮＮ（Ｃｏｎｖｏｌｕｔｉｏｎ　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）の全体コンボリューションレイヤーに対する計算をマッピングするＡＰＩ類似機能を含む。更に、ＤＰＵライブラリ９２１は、ＤＰＵ上へのコンボリューションレイヤー計算のマッピングを最適化するためのＡＰＩ類似機能を含む。

【００５１】

また、ＤＰＵライブラリ９２１は、タスク（例えば、バッチ（ｂａｔｃｈ）、出力チャンネル、ピクセル、入力チャンネル、コンボリューションカーネル）内の全ての個別又は複数の並列性（ｐａｒａｌｌｅｌｉｓｍ）を、チップ、バンク、サブアレイ、及び／又はマットレベルで該当ＤＰＵ並列性にマッピングして、リソース割当を最適化するためのＡＰＩ類似機能を含む。更に、ＤＰＵライブラリ９２１は、性能（即ち、データ移動フロー）と電力消耗との間で均衡を維持（ｔｒａｄｅ　ｏｆｆ）する初期化及び／又はランタイム時に最適のＤＰＵ構成を提供するＡＰＩ類似機能を含む。ＤＰＵライブラリ９２１によって提供される他のＡＰＩ類似機能はデザインノブ（ｋｎｏｂ）タイプ機能を含む。例えば、デザインノブタイプ機能は、バンク当たり活性化されたサブアレイの数、活性化されたサブアレイ当たりの入力機能マップの数、機能マップのパーティショニング（ｐａｒｔｉｔｉｏｎｉｎｇ）、及び／又はコンボリューションカーネルの再使用スキームの設定を含む。また他のＡＰＩ類似機能は、各サブアレイに対して特定タスク（例えば、コンボリューションコンピューティング、チャンネル圧縮（ｓｕｍ　ｕｐ）、及び／又はデータディスパッチング（ｄｉｓｐａｔｃｈｉｎｇ））を割当することによって、追加的なリソースの割当最適化を提供する。オペランドが整数と確率的数字との間で変換された場合、ＤＰＵライブラリ９２１は、精密度制限を満足させながらも、オーバーヘッドを最小化するＡＰＩ類似機能を含む。精密度が予想より低い場合、ＤＰＵライブラリ９２１は、確率的表現のための追加的なビットを使用して値を再び計算するか、又は他のハードウェア（例えば、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ））にタスクを分担（ｏｆｆｌｏａｄ）するＡＰＩ類似機能を含む。

【００５２】

ＤＰＵライブラリ９２１は、またＤＰＵ内の活性化されたサブアレイをスケジュールすると同時にデータ移動をスケジュールして、データ移動がコンピューティング動作によって隠されるＡＰＩ類似機能を含む。

【００５３】

ＤＰＵライブラリ９２１の他の様相は追加的なＤＰＵ開発のための拡張インターフェイスを含む。一実施形態で、ＤＰＵライブラリ９２１は、ＮＯＲ及びシフトロジックを利用して直接機能をプログラムして標準タイプ動作（例えば、加算、乗算、ＭＡＸ／ＭＩＮ等）及び他の動作が提供されるインターフェイスを提供する。拡張インターフェイスは、またインターフェイスを提供し、従ってＤＰＵライブラリ９２１によって具体的に支援されない動作が、ライブラリ及びドライバーレイヤー９２０で、ＳｏＣコントローラ（図示せず）、ＣＰＵ／ＧＰＵ構成要素、及び／又はＣＰＵ／ＴＰＵ構成要素として分担される。ＤＰＵライブラリ９２１の他の様相は、ＤＰＵメモリがコンピューティングのために使用されない場合に、メモリの拡張としてＤＰＵのメモリを使用するためのＡＰＩ類似機能を提供する。

【００５４】

ＤＰＵドライバー９２２は、ＤＰＵハードウェアレイヤーをシステムに集積するために、ハードウェアレイヤー９１０でのＤＰＵ、ＤＰＵライブラリ９２１、及びより高いレイヤーでのオペレーティングシステム（ＯＳ：Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）との間のインターフェイス連結を提供するように構成される。即ち、ＤＰＵドライバー９２２はＤＰＵをシステムＯＳ及びＤＰＵライブラリ９２１に露出する。一実施形態で、ＤＰＵドライバー９２２は初期化時にＤＰＵコントロールを提供する。一実施形態で、ＤＰＵドライバー９２２はＤＲＡＭタイプのアドレス又はＤＲＡＭタイプのアドレスのシーケンスの形態で命令をＤＰＵに伝送し、ＤＰＵの内外へのデータ移動を制御する。ＤＰＵドライバ

ー９２２は、ＤＰＵ－ＣＰＵ及び／又はＤＰＵ－ＧＰＵ通信を処理すると共に多重ＤＰＵ通信を提供する。

【００５５】

ＤＰＵコンパイラ９２３は、ＤＰＵライブラリ９２１からのＤＰＵコードを、ＤＰＵを制御するためにＤＰＵドライバー９２２によって使用されるメモリアドレスの形態であるＤＰＵ命令にコンパイルする。ＤＰＵコンパイラ９２３によって生成されたＤＰＵ命令は、ＤＰＵ内の１つ及び／又は２つのロー上で作動する単一命令（例えば、ベクトル命令、及び／又は集合（ｇａｔｈｅｒｅｄ）ベクトル、リードオン動作命令）である。

【００５６】

フレームワークレイヤー９３０は使いやすい（ｕｓｅｒ－ｆｒｉｅｎｄｌｙ）インターフェイスをライブラリ及びドライバーレイヤー９２０並びにハードウェアレイヤー９１０に提供するように構成される。一実施形態で、フレームワークレイヤー９３０は、アプリケーションレイヤー９４０で広範囲なアプリケーションと互換可能な使いやすいインターフェイスを提供し、ＤＰＵハードウェアレイヤー９１０をユーザーに透過的に（ｔｒａｎｓｐａｒｅｎｔ）提供する。他の実施形態で、フレームワークレイヤー９３０は、定量化（ｑｕａｎｔｉｔａｔｉｏｎ）機能を、例えばＴｏｒｃｈ７タイプアプリケーションやＴｅｎｓｏｒＦｌｏｗタイプアプリケーションのような、既存の、従来の方法に追加するフレームワーク拡張を含む。但し、本発明はこれに制限されない。一実施形態で、フレームワークレイヤー９３０はトレーニングアルゴリズムに定量化機能を追加することを含む。他の実施形態で、フレームワークレイヤー９３０は、既存の割り算、乗算、平方根のバッチ正規化方法に対して、割り算、乗算、平方根の近似方法にシフトする置換を提供する。他の実施形態で、フレームワークレイヤー９３０は、ユーザーが計算のために使用するビット数を設定する拡張を提供する。他の実施形態で、フレームワークレイヤー９３０は、ＤＰＵライブラリ及びドライバーレイヤー９２０からフレームワークレイヤー９３０に多重ＤＰＵ　ＡＰＩをラップ（ｗｒａｐ）するための容量を提供する。従って、ユーザーは複数のＧＰＵの使用と同様にハードウェアレイヤーで複数のＤＰＵを使用することができる。フレームワークレイヤー９３０の他の機能は、ユーザーがハードウェアレイヤー９１０でＤＰＵ又はＧＰＵに機能を割当する。

【００５７】

アプリケーション９４０は、広範囲なアプリケーション（例えば、イメージタグ（ｔａｇ）プロセシング、セルフドライビング／パイロッティング（ｐｉｌｏｔｉｎｇ）運送手段、アルファ碁タイプディープマインドアプリケーション、及び／又は音声研究（ｓｐｅｅｃｈ　ｒｅｓｅａｒｃｈ）等）を含む。但し、本発明はこれに制限されない。

【００５８】

以上、本発明の実施形態について図面を参照しながら詳細に説明したが、本発明は、上述の実施形態に限定されるものではなく、本発明の技術的範囲から逸脱しない範囲内で多様に変更実施することが可能である。

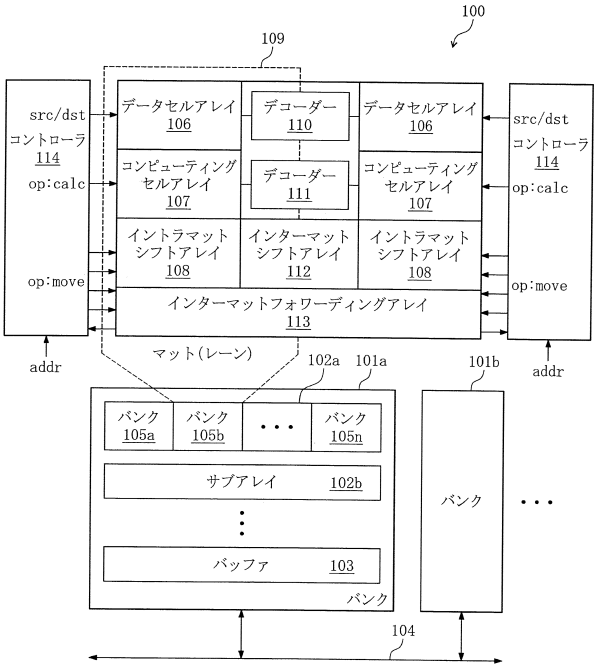
【符号の説明】

【００５９】

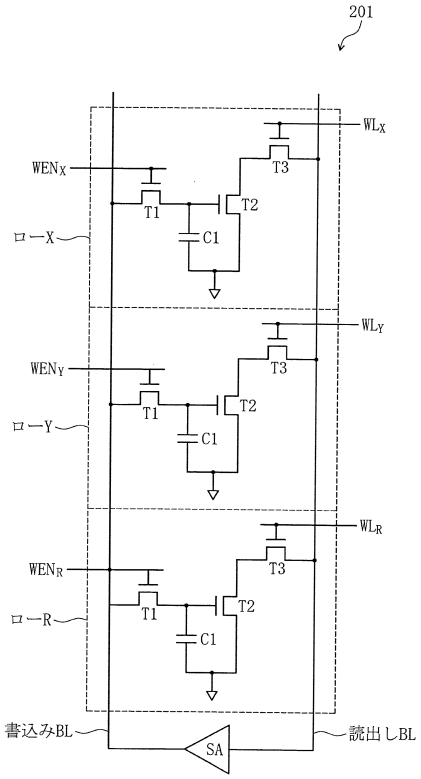
１００、７００　ＤＰＵ
１０１ａ～１０１ｍ、１０５ａ～１０５ｎ　バンク
１０２ａ～１０２ｎ　サブアレイ
１０３　バッファ
１０４　システムバス
１０５、１０５ａ～１０５ｎ、１０９　マット（レーン）
１０６　データセルアレイ
１０７　コンピューティングセルアレイ
１０７ａ～１０７ｄ　コンピューティングセル
１０８　イントラマットシフトアレイ
１１０　データセルアレイデコーダー

１１１　コンピューティングセルアレイデコーダー
１１２　インターマットシフトアレイ
１１２ｅ、１１２ｆ　データシフトライン
１１３　インターマットフォワーディングアレイ
１１３ｇ　第１データフォワーディングライン
１１３ｈ　第２データフォワーディングライン
１１４　サブアレイコントローラ
２０１、２０２　ＤＲＡＭコンピューティングセルトポグラフィ
７１５　確率的データアレイ
７１６　コンバーターツー確率アレイ
９００　システム構造
９１０　ハードウェアレイヤー
９１１　ＰＣＩｅ
９１２　ＤＩＭＭ
９２０　ライブラリ及びドライバーレイヤー
９２１　ＤＰＵライブラリ
９２２　ＤＰＵドライバー
９２３　ＤＰＵコンパイラ
９３０　フレームワークレイヤー
９４０　アプリケーションレイヤー

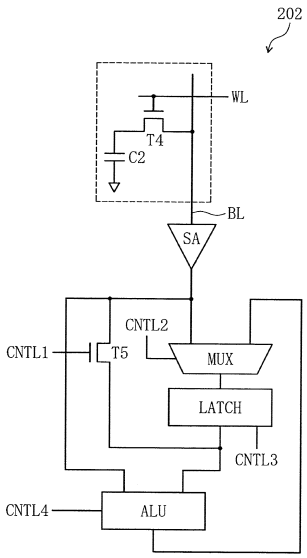
【図１】



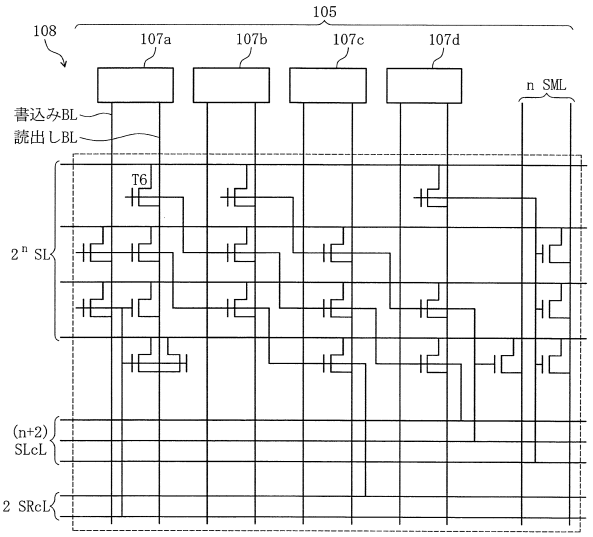
【図２Ａ】



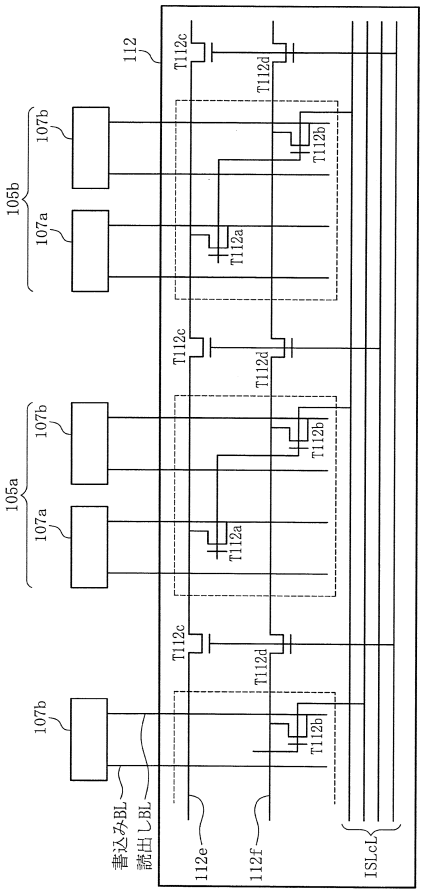
【図２Ｂ】



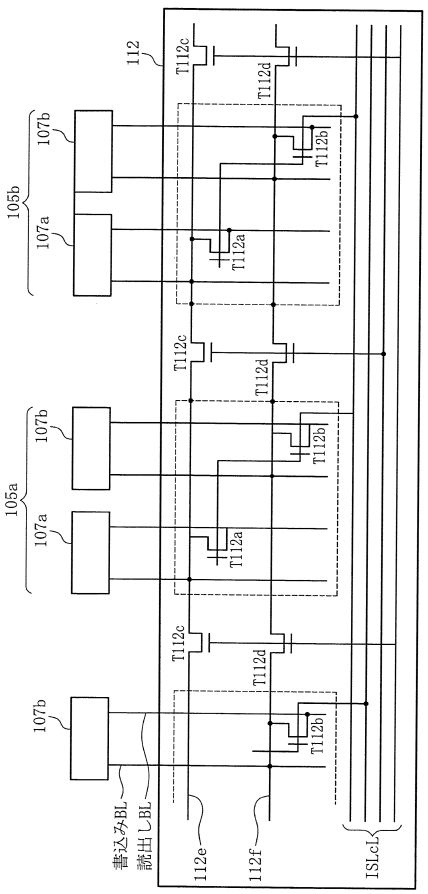
【図３】



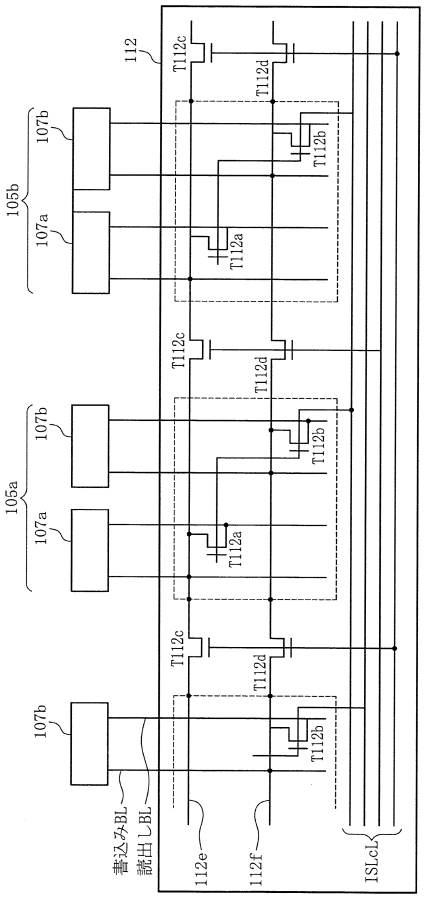
【図４Ａ】



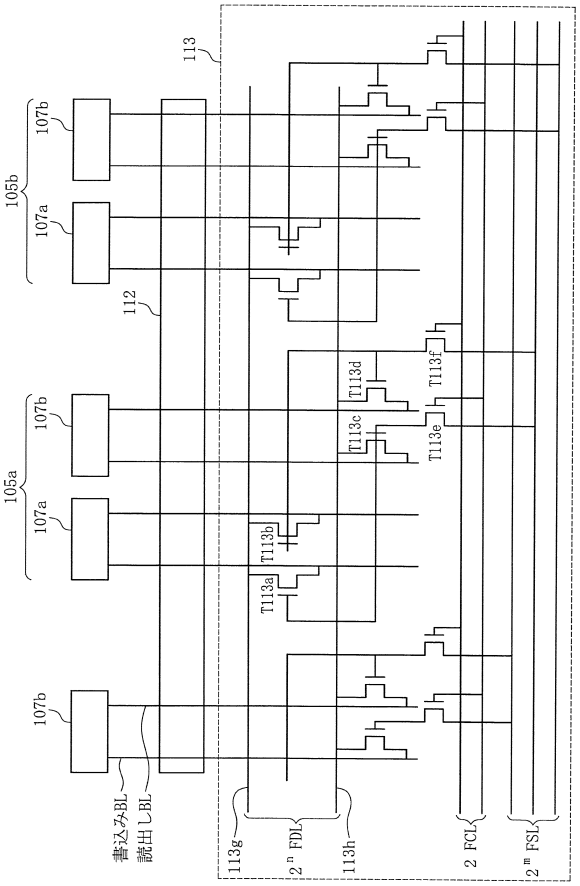
【図４Ｂ】



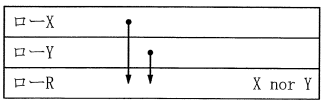
【図４Ｃ】



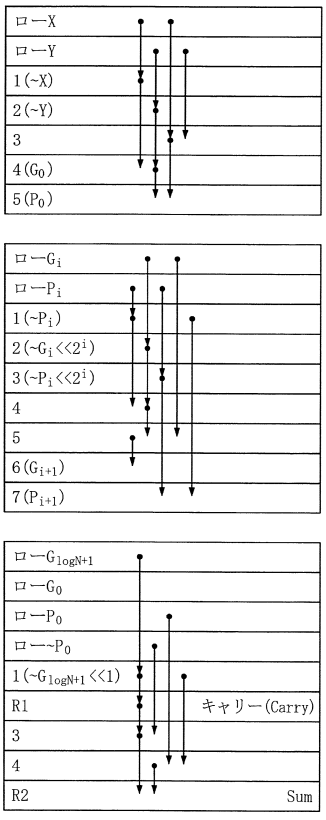
【図５】



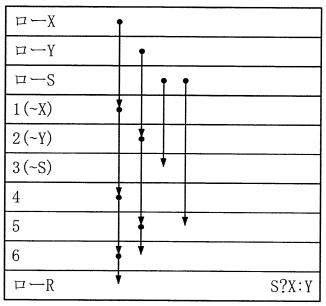
【図６Ａ】



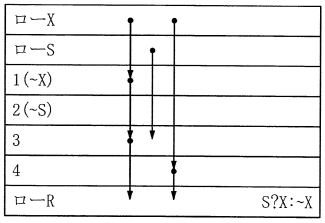
【図６Ｂ】



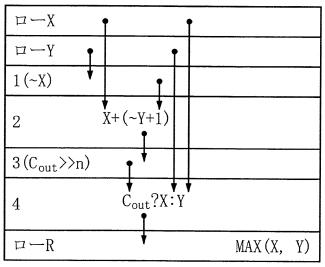
【図６Ｃ】



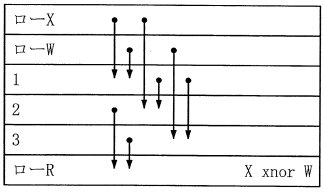
【図６Ｄ】



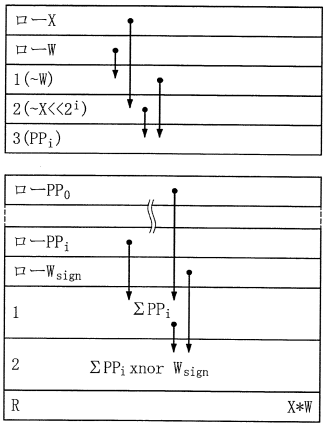
【図６Ｅ】



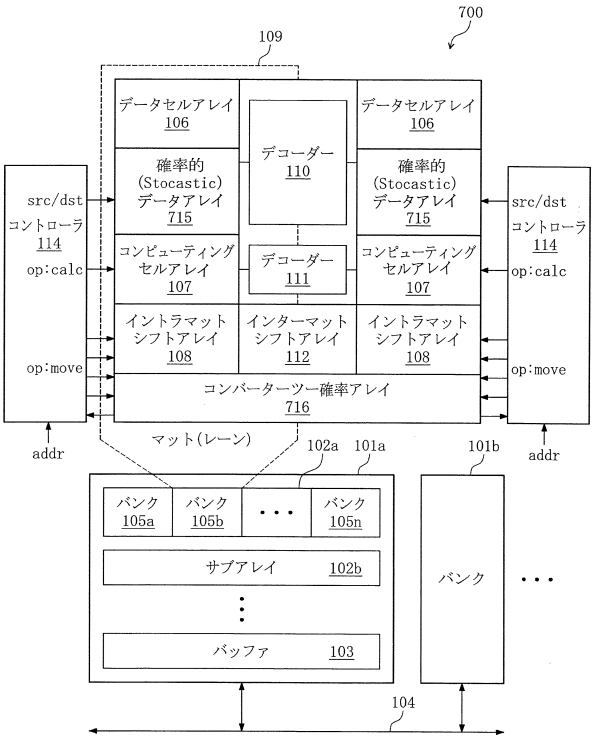
【図６Ｆ】



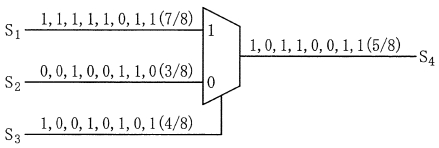
【図６Ｇ】



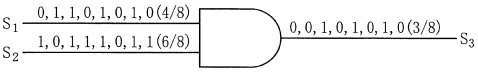
【図７】



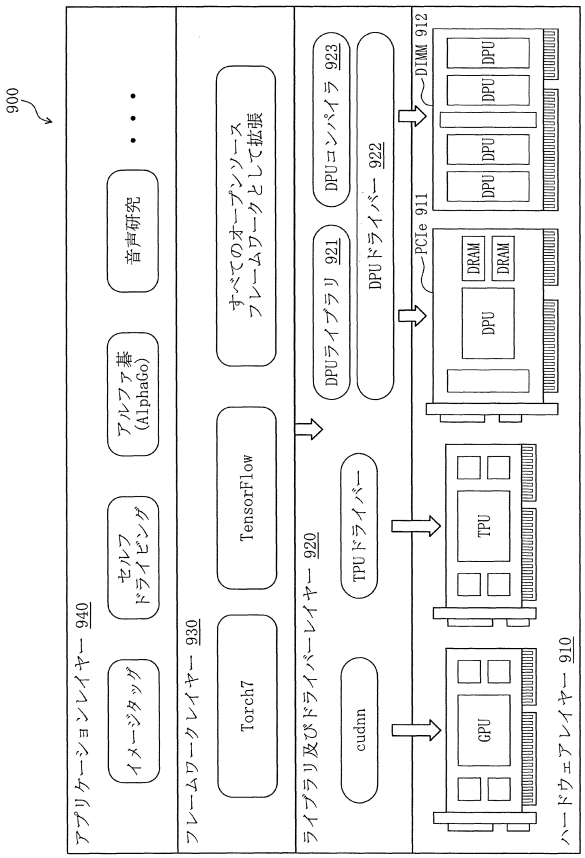
【図８Ａ】



【図８Ｂ】



【図９】


900
アプリケーションレイヤー 940
イメージタッグ
セルフドライビング
アルファ碁 (AlphaGo)
音声研究
・・・
フレームワークレイヤー 930
Torch7
TensorFlow
すべてのオープンソースフレームワークとして拡張
ライブラリ及びドライバーレイヤー 920
cudnn
TPUドライバー
DPUライブラリ 921
DPUドライバー 922
DPUコンパイラ 923
ハードウェアレイヤー 910
GPU
TPU
DPU
DRAM
PCIe 911
DIMM 912

フロントページの続き

(72)発明者 牛 迪 民
アメリカ合衆国, ９４０８７, カリフォルニア州, サニーベール, ホルトハウス テラス ５２７
(72)発明者 マラディ, クリシュナ
アメリカ合衆国, ９５１３５, カリフォルニア州, サン ノゼ, ロートレック ドライブ ４１９６
(72)発明者 鄭 宏 忠
アメリカ合衆国, ９５０３２, カリフォルニア州, ロス ガトス, カールトン アベニュー ＃６, １２０

審査官 後藤 彰

(56)参考文献 特表２０１６－５３６７３３（ＪＰ，Ａ）
特表２０１６－５３２９１９（ＪＰ，Ａ）
米国特許出願公開第２００９／０１６４７８９（ＵＳ，Ａ１）
米国特許第４０６８３０５（ＵＳ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １２／００