



(19)中華民國智慧財產局

(12)發明說明書公告本

(11)證書號數：TW I885266 B

(45)公告日：中華民國 114 (2025) 年 06 月 01 日

(21)申請案號：111117950

(22)申請日：中華民國 111 (2022) 年 05 月 13 日

(51)Int. Cl. : G06N3/063 (2023.01)

G11C11/54 (2006.01)

(30)優先權：2021/05/18 美國

63/190,228

2021/08/30 美國

17/461,901

2021/09/02 世界智慧財產權組織

PCT/US21/48864

(71)申請人：美商超捷公司 (美國) SILICON STORAGE TECHNOLOGY, INC. (US)

美國

(72)發明人：陳 曉萬 TRAN, HIEU VAN (US)；武 順 VU, THUAN (US)；洪 史丹利 HONG,

STANLEY (US)；鄭 史蒂芬 TRINH, STEPHEN (US)；李 英 LY, ANH (US)

(74)代理人：賴經臣；宿希成

(56)參考文獻：

TW 202016928A

US 20170337466A1

US 20200233482A1

WO 2021028723A2

審查人員：林育弘

申請專利範圍項數：1 項 圖式數：39 共 90 頁

(54)名稱

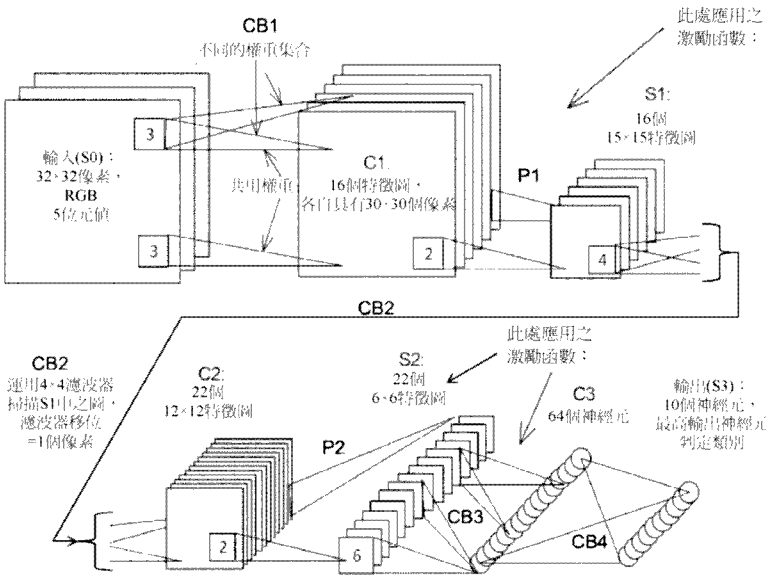
深度學習人工神經網路中類比神經記憶體的分離陣列架構

(57)摘要

本發明揭示用於將一深度學習人工神經網路中之一類比神經記憶體中的一陣列的非揮發性記憶體單元分離成多個部分的眾多具體例。該陣列之每一部分與專用於彼部分之某電路系統以及與該陣列之一或多個其他部分共用之其他電路系統互動。

Numerous embodiments are disclosed for splitting an array of non-volatile memory cells in an analog neural memory in a deep learning artificial neural network into multiple parts. Each part of the array interacts with certain circuitry dedicated to that part and with other circuitry that is shared with one or more other parts of the array.

指定代表圖：



【圖6】

符號簡單說明：

C1,C2,C3,S1,S2:層

CB1,CB2,CB3,CB4:突觸

P1,P2:激勵函數/池化函數

S0:輸入層

S3:輸出層



I885266

【發明摘要】

【中文發明名稱】 深度學習人工神經網路中類比神經記憶體的分離陣列架構

【英文發明名稱】 SPLIT ARRAY ARCHITECTURE FOR ANALOG NEURAL MEMORY IN A DEEP LEARNING ARTIFICIAL NEURAL NETWORK

【中文】

本發明揭示用於將一深度學習人工神經網路中之一類比神經記憶體中的一陣列的非揮發性記憶體單元分離成多個部分的眾多具體例。該陣列之每一部分與專用於彼部分之某電路系統以及與該陣列之一或多個其他部分共用之其他電路系統互動。

【英文】

Numerous embodiments are disclosed for splitting an array of non-volatile memory cells in an analog neural memory in a deep learning artificial neural network into multiple parts. Each part of the array interacts with certain circuitry dedicated to that part and with other circuitry that is shared with one or more other parts of the array.

【指定代表圖】 圖6

【代表圖之符號簡單說明】

C1,C2,C3,S1,S2:層

CB1,CB2,CB3,CB4:突觸

P1,P2:激勵函數/池化函數

S0:輸入層

S3:輸出層

【發明說明書】

【中文發明名稱】 深度學習人工神經網路中類比神經記憶體的分離陣列架構

【英文發明名稱】 SPLIT ARRAY ARCHITECTURE FOR ANALOG NEURAL MEMORY IN A DEEP LEARNING ARTIFICIAL NEURAL NETWORK

【技術領域】

【0001】 [優先權主張]本申請案主張2021年5月18日申請且標題為「深度學習人工神經網路中類比神經記憶體的分離陣列架構(Split Array Architecture for Analog Neural Memory in a Deep Learning Artificial Neural Network)」之美國臨時專利申請案第63/190,228號及2021年8月30日申請且標題為「深度學習人工神經網路中類比神經記憶體的分離陣列架構(Split Array Architecture for Analog Neural Memory in a Deep Learning Artificial Neural Network)」之美國專利申請案第17/461,901號的優先權，該等申請案係以引用方式併入本文中。

【0002】 揭示用於將深度學習人工神經網路中之類比神經記憶體中之一陣列分離成多個部分的眾多具體例，其中每一部分與專用於彼部分之某電路系統及與一或多個其他部分共用之其他電路系統互動。

【先前技術】

【0003】 人工神經網路模擬生物神經網路(動物之中樞神經系統，特定言之，大腦)且用以估計或估算可取決於大量輸入且通常係未知的功能。人工神經網路通常包括彼此交換訊息之多層互連「神經元」。

【0004】 圖1例示人工神經網路，其中圓圈表示神經元之輸入或層。連接(稱為突觸)由箭頭表示，且具有可基於經驗進行調諧之數值權重。此使得神經網路適應於輸入且能夠學習。通常，神經網路包括多個輸入之層。通常存在一或多個中間神經元層及提供神經網路之輸出的輸出神經元層。每一層級處之神經元基於自突觸所接收之資料而個別地或集體地作出決策。

【0005】 用於高效能資訊處理之人工神經網路之發展中的主要挑戰中之一者在於缺乏適當的硬體技術。實際上，切實可行的神經網路依賴於極大數目個突觸，從而實現神經元之間的高連接性，亦即，極高計算並行性。原則上，此複雜性可運用數位超級電腦或特殊化圖形處理單元叢集來達成。然而，除了高成本以外，與生物網路相比，此等方法亦受中等能效困擾，主要因為生物網路執行低精度類比計算，所以其消耗少得多的能量。CMOS類比電路已用於人工神經網路，但鑒於大量神經元及突觸，大部分CMOS實施之突觸已過於龐大。

【0006】 申請人先前在以引用方式併入之美國專利申請案第15/594,439號中揭示了利用一或多個非揮發性記憶體陣列作為突觸之人工(類比)神經網路。非揮發性記憶體陣列作為類比神經形態記憶體進行操作。神經網路構件包括：第一複數個突觸，其經組構以接收第一複數個輸入且自該第一複數個輸入產生第一複數個輸出；及第一複數個神經元，

其經組構以接收該第一複數個輸出。第一複數個突觸包括複數個記憶體單元，其中該等記憶體單元中之每一者包括：形成於半導體基板中之間隔開的源極區及汲極區，其中通道區在源極區與汲極區之間延伸；浮動閘極，其裝設於通道區之第一部分上方且與該第一部分絕緣；以及非浮動閘極，其裝設於通道區之第二部分上方且與該第二部分絕緣。該複數個記憶體單元中之每一者經組構以儲存對應於浮動閘極上之電子數目的權重值。該複數個記憶體單元經組構以使第一複數個輸入乘以經儲存權重值以產生第一複數個輸出。

非揮發性記憶體單元

【0007】 非揮發性記憶體係熟知的。舉例而言，以引用方式併入本文中之美國專利5,029,130 (「'130專利」)揭示了一種分離閘極非揮發性記憶體單元陣列，其為一種類型之快閃記憶體單元。此記憶體單元210顯示於圖2中。每一記憶體單元210包括形成於半導體基板12中之源極區14及汲極區16，其中通道區18處於該源極區與該汲極區之間。浮動閘極20形成於通道區18之第一部分上方並與該第一部分絕緣(且控制該第一部分之導電性)，且形成於源極區14之一部分上方。字線端子22 (其通常耦接至字線)具有：第一部分，其裝設於通道區18之第二部分上方且與該第二部分絕緣(且控制該第二部分之導電性)；及第二部分，其在浮動閘極20上及上方延伸。浮動閘極20及字線端子22藉由閘極氧化物與基板12絕緣。位元線24耦接至汲極區16。

【0008】 記憶體單元210藉由將高正電壓置放於字線端子22上來抹除(其中電子自浮動閘極移除)，此使得浮動閘極20上之電子經由富爾-

諾罕(Fowler-Nordheim；FN)穿隧自浮動閘極20通過中間絕緣件穿隧至字線端子22。

【0009】 記憶體單元210係藉由將正電壓置放於字線端子22上並將正電壓置放於源極區14上而藉由運用熱電子之源極側注入(SSI)而經程式化(其中電子置放於浮動閘極上)。電子電流將自汲極區16朝向源極區14流動。當電子到達字線端子22與浮動閘極20之間的間隙時，該等電子將加速並且變得加熱。經加熱電子中的一些將由於來自浮動閘極20之吸引靜電力而通過閘極氧化物注入至浮動閘極20上。

【0010】 記憶體單元210係藉由將正讀取電壓置於汲極區16及字線端子22上而讀取(此接通在字線端子下方的通道區18之部分)。若浮動閘極20帶正電(亦即，電子經抹除)，則在浮動閘極20下方的通道區18之部分亦經接通，且電流將跨越通道區18流動，此被感測為經抹除或「1」狀態。若浮動閘極20帶負電(亦即，用電子程式化)，則在浮動閘極20下方的通道區之部分被大部分或完全關斷，且電流將不跨越通道區18流動(或將有極少電流跨越通道區18流動)，此被感測為經程式化或「0」狀態。

【0011】 表1描述可施加至記憶體單元110之端子以用於執行讀取、抹除及程式化操作的典型電壓及電流範圍：

表1：圖3之快閃記憶體單元210之操作

	WL	BL	SL
讀取	2-3 V	0.6-2 V	0 V
抹除	~11-13 V	0 V	0 V
程式化	1-2 V	10.5-3 μ A	9-10 V

【0012】 其他分離閘極記憶體單元組構係已知的，該等分離閘極記憶體單元組構係其他類型之快閃記憶體單元。舉例而言，圖3描述四閘極記憶體單元310，其包含源極區14、汲極區16、通道區18之第一部分上方之浮動閘極20、通道區18之第二部分上方之選擇閘極22（通常耦接至字線WL）、浮動閘極20上方之控制閘極28，及源極區14上方之抹除閘極30。此組構描繪於美國專利6,747,310中，該專利出於所有目的以引用方式併入本文中。此處，除浮動閘極20之外的所有閘極皆為非浮動閘極，此意謂該等閘極電連接至或可電連接至電壓源。程式化係藉由來自通道區18之經加熱電子將自身注入至浮動閘極20上予以執行。抹除係藉由自浮動閘極20至抹除閘極30之電子穿隧來執行。

【0013】 表2描述可施加至記憶體單元310之端子以用於執行讀取、抹除及程式化操作的典型電壓及電流範圍：

表2：圖3之快閃記憶體單元310之操作

	WL/SG	BL	CG	EG	SL
讀取	1.0-2V	0.6-2V	0-2.6V	0-2.6V	0V
抹除	-0.5V/0V	0V	0V/-8V	8-12V	0V
程式化	1V	0.1-1μA	8-11V	4.5-9V	4.5-5V

【0014】 圖4描述三閘極記憶體單元410，其為另一類型之快閃記憶體單元。記憶體單元410與圖3之記憶體單元310相同，其例外之處在於記憶體單元410不具有單獨控制閘極。抹除操作(其中抹除經由使用抹除閘極來進行)及讀取操作類似於圖3之抹除操作及讀取操作，惟未施加控制閘極偏壓除外。程式化操作亦在無控制閘極偏壓之情況下進行，且結果，較高電壓在程式化操作期間必須經施加於源極線上以補償控制閘極偏壓之缺乏。

【0015】 表3描述可施加至記憶體單元410之端子以用於執行讀取、抹除及程式化操作的典型電壓及電流範圍：

表3：圖4之快閃記憶體單元410之操作

	WL/SG	BL	EG	SL
讀取	0.7-2.2V	0.6-2V	0-2.6V	0V
抹除	-0.5V/0V	0V	11.5V	0V
程式化	1V	0.2-3μA	4.5V	7-9V

【0016】 圖5描述堆疊閘極記憶體單元510，其為另一類型之快閃記憶體單元。記憶體單元510類似於圖2之記憶體單元210，其例外之處在於浮動閘極20在整個通道區18上方延伸，且控制閘極22（其在此處將耦接至字線）在浮動閘極20上方延伸，藉由絕緣層（圖中未示）分離。抹除係藉由電子自FG至基板之FN穿隧而進行，程式化係藉由通道區18與汲極區16之間的區處進行通道熱電子(CHE)注入、藉由電子自源極區14朝向汲極區16流動來進行，且讀取操作類似於針對具有較高控制閘極電壓之記憶體單元210之讀取操作。

【0017】 表4描述可施加至記憶體單元510之端子以及基板12以用於執行讀取、抹除及程式化操作的典型電壓範圍：

表4：圖5之快閃記憶體單元510之操作

	CG	BL	SL	基板
讀取	2-5V	0.6 - 2V	0V	0V
抹除	-8至-10V/0V	FLT	FLT	8-10V / 15-20V
程式化	8-12V	3-5V	0V	0V

【0018】 本文中所描繪之方法及手段可應用於其他非揮發性記憶體技術，諸如但不限於FINFET分離閘極快閃或堆疊閘極快閃記憶體、NAND快閃、矽氧化氮氧化矽(silicon-oxide-nitride-oxide-silicon；SONOS，氮化物中收集之電荷)、金屬-氧化物-氮化物-氧化物-矽(metal-

oxide-nitride-oxide-silicon；MONOS，氮化物中收集之金屬電荷)、電阻式ram (resistive ram；ReRAM)、相變記憶體(phase change memory；PCM)、磁性ram (magnetic ram；MRAM)、鐵電ram (ferroelectric ram；FeRAM)、電荷收集(charge trap；CT)記憶體、碳管(carbon-tube；CN)記憶體、雙層級或多層級一次性可程式化(bi-level or multi-level one time programmable；OTP)及相關電子ram (correlated electron ram；CeRAM)。

【0019】 為了利用包含上文在人工神經網路中所描繪之非揮發性記憶體單元類型中之一者的記憶體陣列，進行兩種修改。首先，線經結構以使得每一記憶體單元可個別地經程式化、抹除及讀取而不會不利地影響陣列中之其他記憶體單元的記憶體狀態，如下文進一步解釋。其次，記憶體單元之連續(類比)程式化經提供。

【0020】 特定言之，陣列中之每一記憶體單元之記憶體狀態(亦即，浮動閘極上之電荷)可連續地自完全抹除狀態改變為完全程式化狀態，其方式為獨立的且對其他記憶體單元之干擾最小。在另一具體例中，陣列中之每一記憶體單元之記憶體狀態(亦即，浮動閘極上之電荷)可連續地自完全程式化狀態改變為完全抹除狀態，且反之亦然，其方式為獨立的且對其他記憶體單元之干擾最小。此意謂單元儲存係類比的，或至少可儲存許多離散值(諸如16或64個不同值)中之一者，此允許對記憶體陣列中之所有單元進行極精確且個別的調諧，且此使得記憶體陣列對於儲存神經網路之突觸權重及對該等突觸權重進行微調調整係理想的。

採用非揮發性記憶體單元陣列之神經網路

【0021】 圖6在概念上例示利用本發明具體例之非揮發性記憶體陣列的神經網路之非限制性實施例。此實施例將非揮發性記憶體陣列神經網路用於面部辨識應用，但任何其他適當應用皆可使用基於非揮發性記憶體陣列之神經網路來實施。

【0022】 S_0 為輸入層，對於此實施例，該輸入層為具有5位元精度之 32×32 像素RGB影像(亦即，三個 32×32 像素陣列，每種顏色R、G及B一個像素陣列，每一像素為5位元精度)。自輸入層 S_0 行進至層 C_1 之突觸 CB_1 在一些情況下應用不同權重集合且在其他情況下應用共用權重，且用 3×3 像素重疊濾波器(核心)掃描輸入影像，使濾波器移位1個像素(或多於1個像素，如由模型規定)。特定言之，將影像(亦即，被稱作濾波器或核心)之 3×3 部分中之9個像素的值提供至突觸 CB_1 ，其中將此等9個輸入值乘以適當權重，且在對彼乘法之輸出進行求和之後，單個輸出值由第一突觸 CB_1 判定及提供以用於產生層 C_1 之特徵圖中之一者的像素。 3×3 濾波器接著在輸入層 S_0 內向右移位一個像素(亦即，在右側上新增三個像素之行，且在左側上丟棄三個像素之行)，藉以此新定位濾波器中之9個像素值經提供至突觸 CB_1 ，其中使該等像素值乘以相同權重，且第二單一個輸出值係藉由相聯結突觸予以判定。此程序針對所有三種顏色且針對所有位元(精度值)繼續，直至 3×3 濾波器跨越輸入層 S_0 之整個 32×32 像素影像進行掃描為止。程序接著使用不同權重集合進行重複以產生層 C_1 之不同特徵圖，直至層 C_1 之所有特徵圖已經計算為止。

【0023】 在層 C_1 中，在本發明實施例中，存在16個特徵圖，每一特徵圖具有 30×30 個像素。每一像素為自使輸入與核心相乘提取之新特徵

像素，且因此每一特徵圖為二維陣列，且因此在此實施例中，層C1構成二維陣列之16個層(應謹記，本文中所提及之層及陣列為邏輯關係，未必為實體關係，亦即，陣列未必以實體二維陣列定向)。層C1中之16個特徵圖中的每一者皆由應用於濾波器掃描之突觸權重之十六個不同集合中的一者產生。C1特徵圖可皆針對同一影像特徵之不同態樣，諸如邊界識別。舉例而言，第一圖(使用第一權重集合產生，對於用以此第一圖之所有掃描為共用的)可識別圓形邊緣，第二圖(使用不同於第一權重集合之第二權重集合產生)可識別矩形邊緣或某些特徵之縱橫比等等。

【0024】 在自層C1進入層S1之前應用激勵函數P1 (池化)，其池化來自每一特徵圖中之連續非重疊 2×2 區的值。池化函數P1之目的為使附近位置達到平均數(或亦可使用最大函數)，以例如降低邊緣位置之相依性且在進入下一階段之前減小資料大小。在層S1處，存在16個 15×15 特徵圖(亦即，各自具有 15×15 個像素之十六個不同陣列)。自層S1進入層C2之突觸CB2利用 4×4 濾波器掃描S1中之圖，其中濾波器移位1個像素。在層C2處，存在22個 12×12 特徵圖。在自層C2進入層S2之前應用激勵函數P2 (池化)，其池化來自每一特徵圖中之連續非重疊 2×2 區的值。在層S2處，存在22個 6×6 特徵圖。在自層S2進入層C3之突觸CB3處應用激勵函數(池化)，其中層C3中之每一神經元經由CB3之各別突觸連接至層S2中之每個圖。在層C3處，存在64個神經元。自層C3進入輸出層S3之突觸CB4將C3完全連接至S3，亦即，層C3中之每一神經元連接至層S3中之每一神經元。S3處之輸出包括10個神經元，其中最高輸出神經元判定類別。此輸出可例如指示原始影像之內容之識別或分類。

【0025】 每一突觸層係使用非揮發性記憶體單元之陣列或陣列之一部分來實施。

【0026】 圖7為可用於彼目的之陣列的方塊圖。向量乘矩陣乘法(VMM)陣列32包括非揮發性記憶體單元，且用作一層與下一層之間的突觸(諸如圖6中之CB1、CB2、CB3及CB4)。特定言之，VMM陣列32包括一陣列的非揮發性記憶體單元33、抹除閘極及字線閘極解碼器34、控制閘極解碼器35、位元線解碼器36及源極線解碼器37，該等構件對非揮發性記憶體單元陣列33之各別輸入進行解碼。至VMM陣列32之輸入可來自抹除閘極及字線閘極解碼器34或來自控制閘極解碼器35。在此具體例中，源極線解碼器37亦對非揮發性記憶體單元陣列33之輸出進行解碼。替代地，位元線解碼器36可對非揮發性記憶體單元陣列33之輸出進行解碼。

【0027】 非揮發性記憶體單元陣列33用於兩個目的。首先，其儲存將由VMM陣列32使用之權重。其次，非揮發性記憶體單元陣列33有效地使輸入乘以儲存於非揮發性記憶體單元陣列33中之權重，且按輸出線(源極線或位元線)將結果相加以產生輸出，該輸出將為至下一層之輸入或至最終層之輸入。藉由執行乘法及加法函數，非揮發性記憶體單元陣列33消除對單獨的乘法及加法邏輯電路之需求，且由於其原位記憶體計算而亦係功率高效的。

【0028】 將非揮發性記憶體單元陣列33之輸出供應至差分求和器(諸如求和運算放大器或求和電流鏡) 38，該差分求和器對非揮發性記憶體單元陣列33之輸出求和以產生用於彼卷積之單一值。差分求和器38經配置以執行正權重與負權重之求和。

【0029】 接著將差分求和器38之總計輸出值供應至激勵函數區塊39，該激勵函數區塊對輸出進行整流。激勵函數區塊39可提供S型(sigmoid)、雙曲正切(tanh)或ReLU函數。激勵函數區塊39之經整流之輸出值變成作為下一層(例如圖6中之C1)之特徵圖之元素，且接著應用於下一突觸以產生下一特徵圖層或最終層。因此，在此實施例中，非揮發性記憶體單元陣列33構成複數個突觸(其自前一神經元層或自諸如影像資料庫之輸入層接收該等突觸之輸入)，且求和運算放大器38及激勵函數區塊39構成複數個神經元。

【0030】 至圖7中之VMM陣列32之輸入(WLx，EGx，CGx，以及選擇地BLx及SLx)可為類比層級、二進位層級或數位位元(在此情況下，DAC經提供以將數位位元轉換成適當的輸入類比層級)，且輸出可為類比層級、二進位層級或數位位元(在此情況下，輸出ADC經提供以將輸出類比層級轉換成數位位元)。

【0031】 圖8為描述VMM陣列32(此處標記為VMM陣列32a、32b、32c、32d及32e)之眾多層之使用的方塊圖。如圖8中所顯示，表示為Inputx之輸入由數位至類比轉換器31自數位轉換成類比，且經提供至輸入VMM陣列32a。經轉換之類比輸入可為電壓或電流。第一層之輸入D/A轉換可藉由使用函數或查找表(look up table；LUT)來進行，該函數或LUT將輸入Inputx映射至用於輸入VMM陣列32a之矩陣乘法器的適當類比層級。輸入轉換亦可藉由類比至類比(A/A)轉換器來進行以將外部類比輸入轉換成至輸入VMM陣列32a之經映射類比輸入。

【0032】 由輸入VMM陣列32a產生之輸出經提供為至下一VMM陣列(隱藏層級1) 32b之輸入，該下一VMM陣列又產生輸出，該輸出經提供為至下一VMM陣列(隱藏層級2) 32c之輸入，等等。VMM陣列32之各種層充當迴旋神經網路(CNN)之不同的突觸及神經元層。每一VMM陣列32a、32b、32c、32d及32e可為獨立的實體非揮發性記憶體陣列，或多個VMM陣列可利用同一實體非揮發性記憶體陣列之不同部分，或多個VMM陣列可利用同一實體非揮發性記憶體陣列之重疊部分。圖8中所顯示之實施例含有五個層(32a，32b，32c，32d，32e)：一個輸入層(32a)、兩個隱藏層(32b，32c)及兩個完全連接層(32d，32e)。一般熟悉本技藝者將瞭解，此僅為例示性的，且系統替代地可包含多於兩個隱藏層及多於兩個完全連接層。

向量乘矩陣乘法(VMM)陣列

【0033】 圖9描述神經元VMM陣列900，其尤其適合於如圖3中所顯示之記憶體單元310，且用作輸入層與下一層之間的突觸及神經元部分。VMM陣列900包含非揮發性記憶體單元之記憶體陣列901及非揮發性參考記憶體單元之參考陣列902 (在陣列之頂部處)。替代地，另一參考陣列可置放於底部處。

【0034】 在VMM陣列900中，控制閘極線(諸如控制閘極線903)在豎直方向上延行(因此，列方向上之參考陣列902與控制閘極線903正交)，且抹除閘極線(諸如抹除閘極線904)在水平方向上延行。此處，至VMM陣列900之輸入提供於控制閘極線(CG0，CG1，CG2，CG3)上，且VMM陣列900之輸出出現於源極線(SL0，SL1)上。在一個具體例中，僅使用偶數

列，且在另一具體例中，僅使用奇數列。置放於每一源極線(分別為SL0、SL1)上之電流對來自連接至彼特定源極線之記憶體單元的所有電流執行求和函數。

【0035】如本文中針對神經網路所描繪，VMM陣列900之非揮發性記憶體單元，亦即，VMM陣列900之記憶體單元310，較佳經組構以在亞臨限區中操作。

【0036】本文中所描繪之非揮發性參考記憶體單元及非揮發性記憶體單元在弱反轉(亞臨限區)中經偏壓：

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

$$\text{其中 } w = e^{(-V_{th})/nV_t}$$

其中 I_{ds} 為汲極至源極電流； V_g 為記憶體單元上之閘極電壓； V_{th} 為記憶體單元之臨限電壓； V_t 為熱電壓，其 $= k * T / q$ ，其中 k 為波茲曼常數， T 為以克耳文為單位之溫度，且 q 為電子電荷； n 為斜率因數，其 $= 1 + (C_{dep}/C_{ox})$ ，其中 C_{dep} = 耗盡層之電容且 C_{ox} 為閘極氧化物層之電容； I_o 為等於臨限電壓之閘極電壓下之記憶體單元電流， I_o 與 $(W_t/L) * u * C_{ox} * (n-1) * V_t^2$ 成比例，其中 u 為記憶體單元之載流子遷移率，且 W_t 及 L 分別為記憶體單元之寬度及長度。

【0037】對於使用記憶體單元(諸如參考記憶體單元或周邊記憶體單元)或電晶體將輸入電流轉換成輸入電壓之I至V對數轉換器：

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

其中， w_p 為參考或周邊記憶體單元之 w 。

【0038】 對於用作具有電流輸入之向量矩陣乘法器VMM陣列之記憶體陣列，輸出電流為：

$$I_{out} = w_a * I_o * e^{(V_g)/nV_t} , \text{ 即}$$

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

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

此處， w_a = 記憶體陣列中之每一記憶體單元之 w 。

V_{thp} 為周邊記憶體單元之有效臨限電壓，且 V_{tha} 為主(資料)記憶體單元之有效臨限電壓。應注意，電晶體之臨限電壓為基板基底偏壓電壓之函數，且表示為 V_{sb} 之基板基底偏壓電壓可經調變以補償此溫度下的各種條件。臨限電壓 V_{th} 可表示為：

$$V_{th} = V_{th0} + \gamma (\text{SQRT } |V_{sb} - 2 * \phi_F) - \text{SQRT } |2 * \phi_F|)$$

其中 V_{th0} 為具有零基板偏壓之臨限電壓， ϕ_F 為表面電位，且 γ 為體效應參數。

【0039】 字線或控制閘極可用作用於輸入電壓之記憶體單元之輸入。

【0040】 替代地，本文中所描繪之VMM陣列之快閃記憶體單元可經組構以在線性區中操作：

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

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

此意謂線性區中之權重 W 係與 $(V_{gs} - V_{th})$ 成比例。

【0041】 字線或控制閘極或位元線或源極線可用作在線性區中操作之記憶體單元的輸入。位元線或源極線可用作記憶體單元之輸出。

【0042】 對於I至V線性轉換器，記憶體單元(諸如，參考記憶體單元或周邊記憶體單元)或在線性區中操作之電晶體可用以將輸入/輸出電流線性地轉換成輸入/輸出電壓。

【0043】 替代地，本文中所描繪之VMM陣列之記憶體單元可經組構以在飽和區中操作：

$$I_{ds} = \frac{1}{2} * \beta * (V_{gs} - V_{th})^2 ; \beta = \mu * C_{ox} * W_t / L$$

$W \propto (V_{gs} - V_{th})^2$ ，此意謂權重W係與 $(V_{gs} - V_{th})^2$ 成比例。

【0044】 字線、控制閘極或抹除閘極可用作在飽和區中操作之記憶體單元的輸入。位元線或源極線可用作輸出神經元之輸出。

【0045】 替代地，本文中所描繪之VMM陣列之記憶體單元可用於神經網路之每一層或多層之所有區或其組合(次臨限區、線性區或飽和區)中。

【0046】 用於圖7之VMM陣列32的其他具體例描繪於以引用方式併入本文中之美國專利申請案第15/826,345號中。如彼申請案中所描繪，源極線或位元線可用作神經元輸出(電流求和輸出)。

【0047】 圖10描述神經元VMM陣列1000，其尤其適合於如圖2中所顯示之記憶體單元210，且用作輸入層與下一層之間的突觸。VMM陣列1000包含非揮發性記憶體單元之記憶體陣列1003、第一非揮發性參考記憶體單元之參考陣列1001及第二非揮發性參考記憶體單元之參考陣列1002。配置於陣列之行方向上的參考陣列1001及1002用以將流入端子BLR0、BLR1、BLR2及BLR3中之電流輸入轉換成電壓輸入WL0、WL1、WL2及WL3。實際上，第一及第二非揮發性參考記憶體單元為二極體連

接式貫穿多工器1014 (僅部分描述)，其中電流輸入流入該等多工器中。參考單元經調諧(例如，經程式化)至目標參考層級。目標參考層級係由參考小型陣列矩陣(圖中未示)提供。

【0048】 記憶體陣列1003用於兩個目的。首先，其儲存將由VMM陣列1000在其各別記憶體單元上使用之權重。其次，記憶體陣列1003有效地使輸入(亦即，在端子BLR0、BLR1、BLR2及BLR3中提供之電流輸入，參考陣列1001及1002將其轉換成輸入電壓以供應至字線WL0、WL1、WL2及WL3)乘以儲存於記憶體陣列1003中之權重，且隨後將所有結果(記憶體單元電流)相加以在各別位元線(BL0至BLN)上產生輸出，該輸出將為至下一層之輸入或至最終層之輸入。藉由執行乘法及加法函數，記憶體陣列1003消除對單獨的乘法及加法邏輯電路之需求，且亦係功率高效的。此處，電壓輸入提供於字線WL0、WL1、WL2及WL3上，且輸出在讀取(推斷)操作期間出現於各別位元線BL0至BLN上。置於位元線BL0至BLN中之每一者上的電流對來自連接至彼特定位元線之所有非揮發性記憶體單元的電流執行求和函數。

【0049】 表5描述用於VMM陣列1000之操作電壓及電流。該表中之行指示置放於以下各者上的電壓：用於選定單元之字線、用於未選定單元之字線、用於選定單元之位元線、用於未選定單元之位元線、用於選定單元之源極線及用於未選定單元之源極線。列指示讀取、抹除及程式化之操作。

表5：圖10之VMM陣列1000之操作：

	WL	WL-未選定	BL	BL-未選定	SL	SL-未選定
讀取	1-3.5V	-0.5V/0V	0.6-2V (Ineuron)	0.6V-2V/0V	0V	0V
抹除	~5-13V	0V	0V	0V	0V	0V
程式化	1-2V	-0.5V/0V	0.1-3 uA	Vinh ~2.5V	4-10V	0-1V/FLT

【0050】 圖11描述神經元VMM陣列1100，其尤其適合於如圖2中所顯示之記憶體單元210，且用作輸入層與下一層之間的突觸及神經元部分。VMM陣列1100包含非揮發性記憶體單元之記憶體陣列1103、第一非揮發性參考記憶體單元之參考陣列1101及第二非揮發性參考記憶體單元之參考陣列1102。參考陣列1101及1102在VMM陣列1100之列方向上延行。VMM陣列類似於VMM 1000，其例外之處在於在VMM陣列1100中，字線在豎直方向上延行。此處，輸入提供於字線(WLA0，WLB0，WLA1，WLB1，WLA2，WLB2，WLA3，WLB3)上，且輸出在讀取操作期間出現於源極線(SL0，SL1)上。置放於每一源極線上之電流對來自連接至彼特定源極線之記憶體單元的所有電流執行求和函數。

【0051】 表6描述用於VMM陣列1100之操作電壓及電流。該表中之行指示置放於以下各者上的電壓：用於選定單元之字線、用於未選定單元之字線、用於選定單元之位元線、用於未選定單元之位元線、用於選定單元之源極線及用於未選定單元之源極線。列指示讀取、抹除及程式化之操作。

表6：圖11之VMM陣列1100之操作

	WL	WL-未選定	BL	BL-未選定	SL	SL-未選定
讀取	1-3.5V	-0.5V/0V	0.6-2V	0.6V-2V/0V	~0.3-1V (Ineuron)	0V
抹除	~5-13V	0V	0V	0V	0V	SL-抑制(~4- 8V)
程式化	1-2V	-0.5V/0V	0.1-3 uA	Vinh ~2.5V	4-10V	0-1V/FLT

【0052】 圖12描述神經元VMM陣列1200，其尤其適合於如圖3中所顯示之記憶體單元310，且用作輸入層與下一層之間的突觸及神經元部分。VMM陣列1200包含非揮發性記憶體單元之記憶體陣列1203、第一非揮發性參考記憶體單元之參考陣列1201及第二非揮發性參考記憶體單元之參考陣列1202。參考陣列1201及1202用以將流入端子BLR0、BLR1、BLR2及BLR3中之電流輸入轉換成電壓輸入CG0、CG1、CG2及CG3。實際上，第一及第二非揮發性參考記憶體單元為二極體連接式貫穿多工器1212 (僅部分顯示)，其中電流輸入經由BLR0、BLR1、BLR2及BLR3流入該等多工器中。多工器1212各自包括各別多工器1205及串疊電晶體1204以確保在讀取操作期間第一及第二非揮發性參考記憶體單元中之每一者之位元線(諸如BLR0)上的恆定電壓。參考單元經調諧至目標參考層級。

【0053】 記憶體陣列1203用於兩個目的。首先，其儲存將由VMM陣列1200使用之權重。其次，記憶體陣列1203有效地使輸入(提供至端子BLR0、BLR1、BLR2及BLR3之電流輸入，其中參考陣列1201及1202將此等電流輸入轉換成輸入電壓以供應至控制閘極(CG0，CG1，CG2及CG3))乘以儲存於記憶體陣列中之權重，且接著將所有結果(單元電流)相加以產生輸出，該輸出顯現於BL0至BLN上，且將為至下一層之輸入或至最終層之輸入。藉由執行乘法及加法函數，記憶體陣列消除對單獨的乘法及加法邏輯電路之需求，且亦係功率高效的。此處，輸入提供於控制閘極線(CG0，CG1，CG2及CG3)上，且輸出在讀取操作期間出現於位元線

(BL0至BLN)上。置於每一位元線上之電流對來自連接至彼特定位元線之記憶體單元的所有電流執行求和函數。

【0054】 VMM陣列1200針對記憶體陣列1203中之非揮發性記憶體單元實施單向調諧。亦即，每一非揮發性記憶體單元經抹除且接著經部分程式化，直至達到浮動閘極上之所需電荷。若過多電荷被置放於浮動閘極上(使得錯誤值儲存於單元中)，則單元被抹除且部分程式化操作之序列重新開始。如所顯示，共用同一抹除閘極(諸如EG0或EG1)之兩個列被一起抹除(此被稱為頁面抹除)，且此後，每一單元經部分地程式化直至達到浮動閘極上之所需電荷。

【0055】 表7描述用於VMM陣列1200之操作電壓及電流。該表中之行指示置放於以下各者上的電壓：用於選定單元之字線、用於未選定單元之字線、用於選定單元之位元線、用於未選定單元之位元線、用於選定單元之控制閘極、用於與選定單元在同一扇區中之未選定單元的控制閘極、用於與選定單元在不同扇區中之未選定單元的控制閘極、用於選定單元之抹除閘極、用於未選定單元之抹除閘極、用於選定單元之源極線及用於未選定單元之源極線。列指示讀取、抹除及程式化之操作。

表7：圖12之VMM陣列1200之操作

	WL	WL-未選定	BL	BL-未選定	CG	CG-未選定同一扇區	CG-未選定	EG	EG-未選定	SL	SL-未選定
讀取	1.0-2V	-0.5V/ 0V	0.6-2V (Ineuron)	0V	0-2.6V	0-2.6V	0-2.6V	0-2.6V	0-2.6V	0V	0V
抹除	0V	0V	0V	0V	0V	0-2.6V	0-2.6V	5-12V	0-2.6V	0V	0V
程式化	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

【0056】 圖13描述神經元VMM陣列1300，其尤其適合於如圖3中所顯示之記憶體單元310，且用作輸入層與下一層之間的突觸及神經元部分。VMM陣列1300包含非揮發性記憶體單元之記憶體陣列1303、參考陣列1301或第一非揮發性參考記憶體單元及第二非揮發性參考記憶體單元之參考陣列1302。EG線EGR0、EG0、EG1及EGR1豎直地延行，而CG線CG0、CG1、CG2及CG3以及SL線WL0、WL1、WL2及WL3水平地延行。VMM陣列1300類似於VMM陣列1400，其例外之處在於VMM陣列1300實施雙向調諧，其中由於使用單獨的EG線，每一個別單元可視需要經完全抹除、部分程式化及部分抹除以達到浮動閘極上之所需電荷量。如所顯示，參考陣列1301及1302將端子BLR0、BLR1、BLR2及BLR3中之輸入電流轉換成待在列方向上施加至記憶體單元之控制閘極電壓CG0、CG1、CG2及CG3 (經由二極體連接式參考單元貫穿多工器1314進行之動作)。電流輸出(神經元)在位元線BL0至BLN中，其中每一位元線對來自連接至彼特定位元線之非揮發性記憶體單元的所有電流進行求和。

【0057】 表8描述用於VMM陣列1300之操作電壓及電流。該表中之行指示置放於以下各者上的電壓：用於選定單元之字線、用於未選定單元之字線、用於選定單元之位元線、用於未選定單元之位元線、用於選定單元之控制閘極、用於與選定單元在同一扇區中之未選定單元的控制閘極、用於與選定單元在不同扇區中之未選定單元的控制閘極、用於選定單元之抹除閘極、用於未選定單元之抹除閘極、用於選定單元之源極線及用於未選定單元之源極線。列指示讀取、抹除及程式化之操作。

表8：圖13之VMM陣列1300之操作

	WL	WL-未選定	BL	BL-未選定	CG	CG-未選定 同一扇區	CG-未選定	EG	EG-未選定	SL	SL-未選定
讀取	1.0-2V	-0.5 V/ 0V	0.6-2V (Incuron)	0V	0-2.6V	0 -2.6V	0-2.6V	0-2.6V	0-2.6V	0V	0V
抹除	0V	0V	0V	0V	0V	4-9V	0-2.6V	5-12V	0-2.6V	0V	0V
程式化	0.7-1V	-0.5 V/ 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

【0058】 圖22描述神經元VMM陣列2200，其尤其適於如圖2中所示之記憶體單元210，且用作輸入層與下一層之間的突觸及神經元部分。在VMM陣列2200中，輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_N$ 分別在位元線 BL_0 、 $\dots$ 、 BL_N 上經接收，且輸出 $OUTPUT_1$ 、 $OUTPUT_2$ 、 $OUTPUT_3$ 及 $OUTPUT_4$ 分別產生於源極線 SL_0 、 SL_1 、 SL_2 及 SL_3 上。

【0059】 圖23描述神經元VMM陣列2300，其尤其適於如圖2中所示之記憶體單元210，且用作輸入層與下一層之間的突觸及神經元部分。在此實施例中，輸入 $INPUT_0$ 、 $INPUT_1$ 、 $INPUT_2$ 及 $INPUT_3$ 分別在源極線 SL_0 、 SL_1 、 SL_2 及 SL_3 上經接收，且輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 產生於位元線 BL_0 、 $\dots$ 、 BL_N 上。

【0060】 圖24描述神經元VMM陣列2400，其尤其適於如圖2中所示之記憶體單元210，且用作輸入層與下一層之間的突觸及神經元部分。在此實施例中，輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ 分別在字線 WL_0 、 $\dots$ 、 WL_M 上經接收，且輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 產生於位元線 BL_0 、 $\dots$ 、 BL_N 上。

【0061】 圖25描述神經元VMM陣列2500，其尤其適於如圖3中所示之記憶體單元310，且用作輸入層與下一層之間的突觸及神經元部分。

在此實施例中，輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ 分別在字線 WL_0 、 $\dots$ 、 WL_M 上經接收，且輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 產生於位元線 BL_0 、 $\dots$ 、 BL_N 上。替代地，輸入可在控制閘極 CG_0 、 $\dots$ 、 CG_M 上經接收。

【0062】 圖26A描述神經元VMM陣列2600，其尤其適於如圖4中所示之記憶體單元410，且用作輸入層與下一層之間的突觸及神經元部分。在此實施例中，輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_n$ 分別在豎直控制閘極線 CG_0 、 $\dots$ 、 CG_N 上經接收，且輸出 $OUTPUT_1$ 及 $OUTPUT_2$ 產生於源極線 SL_0 及 SL_1 上。

【0063】 圖26B描述神經元VMM陣列2620，其為具有豎直而非水平字線的VMM陣列2600之替代設計。在此情況下，輸入可在豎直字線 WL_0 、 WL_1 上經接收，且輸出 $OUTPUT_1$ 及 $OUTPUT_2$ 產生於水平源極線 SL_0 及 SL_1 上。

【0064】 圖26C描述神經元VMM陣列2640，其為具有豎直而非水平抹除閘極線的VMM陣列2600之替代設計。在此情況下，輸入可在豎直抹除閘極線 EG_0 、 EG_1 上經接收，且輸出 $OUTPUT_1$ 及 $OUTPUT_2$ 產生於水平源極線 SL_0 及 SL_1 上，且輸出 $OUTPUT_1$ 及 $OUTPUT_2$ 產生於水平源極線 SL_0 及 SL_1 上。

【0065】 圖27描述神經元VMM陣列2700，其尤其適於如圖4中所示之記憶體單元410，且用作輸入層與下一層之間的突觸及神經元部分。在此實施例中，輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_N$ 分別在位元線控制閘極2701-1、2701-2、 $\dots$ 、2701-(N-1)及2701-N之閘極上經接收，該等位元線控制

閘極分別耦接至位元線 BL_0 、 $\dots$ 、 BL_N 。例示性輸出 $OUTPUT_1$ 及 $OUTPUT_2$ 產生於源極線 SL_0 及 SL_1 上。

【0066】 圖28描述神經元VMM陣列2800，其尤其適合於如圖3中所顯示之記憶體單元310、如圖5中所顯示之記憶體單元510及如圖7中所顯示之記憶體單元710，且用作輸入層與下一層之間的突觸及神經元部分。在此實施例中，輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ 在字線 WL_0 、 $\dots$ 、 WL_M 上經接收，且輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 分別產生於位元線 BL_0 、 $\dots$ 、 BL_N 上。替代地，輸入可在控制閘極線 CG_0 、 $\dots$ 、 CG_M 上經接收。

【0067】 圖29描述神經元VMM陣列2900，其尤其適合於如圖3中所顯示之記憶體單元310、如圖5中所顯示之記憶體單元510及如圖7中所顯示之記憶體單元710，且用作輸入層與下一層之間的突觸及神經元部分。在此實施例中，輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ 在控制閘極線 CG_0 、 $\dots$ 、 CG_M 上經接收。輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 分別產生於豎直源極線 SL_0 、 $\dots$ 、 SL_N 上，其中每一源極線 SL_i 耦接至行 i 中之所有記憶體單元之源極線。替代地，輸入可在字線 WL_0 、 $\dots$ 、 WL_M 上經接收。

【0068】 圖30描述神經元VMM陣列3000，其尤其適合於如圖3中所顯示之記憶體單元310、如圖5中所顯示之記憶體單元510及如圖7中所顯示之記憶體單元710，且用作輸入層與下一層之間的突觸及神經元部分。在此實施例中，輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ 在控制閘極線 CG_0 、 $\dots$ 、 CG_M 上經接收。輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 分別產生於豎直位元線 BL_0 、 $\dots$ 、 BL_N 上，其中每一位元線 BL_i 耦接至行 i 中之所有記憶體單元之位元線。

長短期記憶體

【0069】 先前技術包括被稱為長短期記憶體(LSTM)之概念。LSTM單元常常用於神經網路。LSTM允許神經網路在預定任意時間間隔內記住資訊且在後續操作中使用彼資訊。習知LSTM單元包含單元(cell)、輸入閘極、輸出閘極及遺忘閘極。三個閘極調節資訊進入及離開單元之流動及在LSTM中記住資訊的時間間隔。VMM尤其適用於LSTM單元。

【0070】 圖14描述例示性LSTM 1400。此實施例中之LSTM 1400包含單元1401、1402、1403及1404。單元1401接收輸入向量 x_0 ，且產生輸出向量 h_0 及單元狀態向量 c_0 。單元1402接收輸入向量 x_1 、來自單元1401之輸出向量(隱藏狀態) h_0 及來自單元1401之單元狀態 c_0 ，且產生輸出向量 h_1 及單元狀態向量 c_1 。單元1403接收輸入向量 x_2 、來自單元1402之輸出向量(隱藏狀態) h_1 及來自單元1402之單元狀態 c_1 ，且產生輸出向量 h_2 及單元狀態向量 c_2 。單元1404接收輸入向量 x_3 、來自單元1403之輸出向量(隱藏狀態) h_2 及來自單元1403之單元狀態 c_2 ，且產生輸出向量 h_3 。可使用額外單元，且具有四個單元之LSTM僅為實施例。

【0071】 圖15描述LSTM單元1500之例示性實施，其可用於圖14中之單元1401、1402、1403及1404。LSTM單元1500接收輸入向量 $x(t)$ 、來自前一單元之單元狀態向量 $c(t-1)$ 及來自前一單元之輸出向量 $h(t-1)$ ，且產生單元狀態向量 $c(t)$ 及輸出向量 $h(t)$ 。

【0072】 LSTM單元1500包含S型函數構件1501、1502及1503，其中之每一者應用0與1之間的數字以控制輸入向量中之每一分量被允許通

過輸出向量之量。LSTM單元1500亦包含用以將雙曲正切函數應用於輸入向量之雙曲正切構件1504及1505、用以使兩個向量相乘在一起之乘法器構件1506、1507及1508，及用以將兩個向量相加在一起之加法構件1509。輸出向量 $h(t)$ 可提供至系統中之下一LSTM單元，或其可經存取以用於其他目的。

【0073】 圖16描述LSTM單元1600，其為LSTM單元1500之實施之實施例。為了方便讀者，來自LSTM單元1500之相同編號用於LSTM單元1600中。S型函數構件1501、1502及1503以及雙曲正切構件1504各自包含多個VMM陣列1601及激勵函數區塊1602。因此，可見VMM陣列尤其適用於在某些神經網路系統中使用之LSTM單元。乘法器構件1506、1507及1508以及加法構件1509以數位方式或以類比方式實施。激勵函數區塊1602可以數位方式或以類比方式實施。

【0074】 LSTM單元1600之替代方案(及LSTM單元1500之實施之另一實施例)在圖17中加以顯示。在圖17中，S型函數構件1501、1502及1503以及雙曲正切構件1504以時間多工方式共用同一實體硬體(VMM陣列1701及激勵函數區塊1702)。LSTM單元1700亦包含用以使兩個向量相乘在一起之乘法器構件1703、用以使兩個向量相加在一起之加法構件1708、雙曲正切構件1505（其包含激勵函數區塊1702）、用以當 $i(t)$ 自S型函數區塊1702輸出時儲存值 $i(t)$ 之暫存器1707、用以當值 $f(t)*c(t-1)$ 通過多工器1710自乘法器構件1703輸出時儲存彼值之暫存器1704、用以當值 $i(t)*u(t)$ 經由多工器1710自乘法器構件1703輸出時儲存彼值之暫存

器1705，及用以當值 $o(t) * c_{\sim}(t)$ 經由多工器1710及多工器1709自乘法器構件1703輸出時儲存彼值之暫存器1706。

【0075】 LSTM單元1600含有VMM陣列1601及各別激勵函數區塊1602之多個集合，而LSTM單元1700僅含有VMM陣列1701及激勵函數區塊1702之一個集合，該等VMM陣列1701及該激勵函數區塊1702用於表示LSTM單元1700之具體例中之多個層。LSTM單元1700將需要相較於LSTM 1600較少的空間，此係因為LSTM單元1700相比於LSTM單元1600將需要1/4之空間用於VMM及激勵函數區塊。

【0076】 可進一步瞭解，LSTM單元將通常包含多個VMM陣列，其中之每一者需要由VMM陣列外部之某些電路區塊(諸如求和器及激勵函數區塊以及高電壓產生區塊)提供的功能性。為每一VMM陣列提供單獨電路區塊將需要半導體構件內之大量空間且將略微低效。因此，下文所描繪之具體例試圖最小化VMM陣列自身外部所需之電路系統。

閘控遞回單元

【0077】 類比VMM實施可用於閘控遞回單元(gated recurrent unit ; GRU)系統。GRU為遞回神經網路中之閘控機構。GRU類似於LSTM，其例外之處在於GRU單元通常含有比LSTM單元少的組件。

【0078】 圖18描述例示性GRU 1800。此實施例中之GRU 1800包含單元1801、1802、1803及1804。單元1801接收輸入向量 x_0 且產生輸出向量 h_0 。單元1802接收輸入向量 x_1 、來自單元1801之輸出向量 h_0 ，且產生輸出向量 h_1 。單元1803接收輸入向量 x_2 及來自單元1802之輸出向量(隱藏狀態) h_1 ，且產生輸出向量 h_2 。單元1804接收輸入向量 x_3 及來自單元

1803之輸出向量(隱藏狀態) h_2 且產生輸出向量 h_3 。可使用額外單元，且具有四個單元之GRU僅為實施例。

【0079】 圖19描述GRU單元1900之例示性實施，其可用於圖18之單元1801、1802、1803及1804。GRU單元1900接收輸入向量 $x(t)$ 及來自前一GRU單元之輸出向量 $h(t-1)$ ，且產生輸出向量 $h(t)$ 。GRU單元1900包含S型函數構件1901及1902，其中之每一者將0與1之間的數字應用至來自輸出向量 $h(t-1)$ 及輸入向量 $x(t)$ 之分量。GRU單元1900亦包含用以將雙曲正切函數應用至輸入向量之雙曲正切構件1903，用以將兩個向量相乘在一起之複數個乘法器構件1904、1905及1906，用以將兩個向量相加在一起之加法構件1907及用以自1減去輸入以產生輸出之互補構件1908。

【0080】 圖20描述GRU單元2000，其為GRU單元1900之實施之實施例。為方便讀者，與GRU單元1900相同之編號用於GRU單元2000中。如圖20中可見，S型函數構件1901及1902以及雙曲正切構件1903各自包含多個VMM陣列2001及激勵函數區塊2002。因此，可見VMM陣列尤其適用於在某些神經網路系統中使用之GRU單元。乘法器構件1904、1905、1906、加法構件1907及互補構件1908以數位方式或以類比方式實施。激勵函數區塊2002可以數位方式或以類比方式實施。

【0081】 GRU單元2000之替代方案(及GRU單元1900之實施之另一實施例)在圖21中加以顯示。在圖21中，GRU單元2100利用VMM陣列2101及激勵函數區塊2102，該激勵函數區塊在經組構為S型函數時應用0與1之間的數字以控制輸入向量中之每一分量被允許通過輸出向量之量。在圖21中，S型函數構件1901及1902以及雙曲正切構件1903以時間多工

方式共用同一實體硬體(VMM陣列2101及激勵函數區塊2102)。GRU單元2100亦包含用以使兩個向量相乘在一起之乘法器構件2103、用以使兩個向量相加在一起之加法構件2105、用以自1減去輸入以產生輸出之互補構件2109、多工器2104、用以當值 $h(t-1)*r(t)$ 經由多工器2104自乘法器構件2103輸出時保存彼值之暫存器2106、用以當值 $h(t-1)*z(t)$ 經由多工器2104自乘法器構件2103輸出時保存彼值之暫存器2107，及用以當值 $h^*(t)*(1-z(t))$ 經由多工器2104自乘法器構件2103輸出時保存彼值之暫存器2108。

【0082】 GRU單元2000含有VMM陣列2001及激勵函數區塊2002之多個集合，而GRU單元2100僅含有VMM陣列2101及激勵函數區塊2102之一個集合，該等VMM陣列及該激勵函數區塊用於表示GRU單元2100之具體例中之多個層。GRU單元2100將需要相較於GRU單元2000較少的空間，此係因為GRU單元2100相比於GRU單元2000將需要1/3之空間用於VMM及激勵函數區塊。

【0083】 可進一步瞭解，GRU系統將通常包含多個VMM陣列，其中之每一者需要由VMM陣列外部之某些電路區塊(諸如求和器及激勵函數區塊以及高電壓產生區塊)提供的功能性。為每一VMM陣列提供單獨電路區塊將需要半導體構件內之大量空間且將略微低效。因此，下文所描繪之具體例試圖最小化VMM陣列自身外部所需之電路系統。

【0084】 至VMM陣列之輸入可為類比層級、二進位層級、脈衝、時間經調變脈衝或數位位元(在此情況下，需要DAC將數位位元轉換成適當的輸入類比層級)，且輸出可為類比層級、二進位層級、定時脈衝、脈衝

或數位位元(在此情況下，需要輸出ADC將輸出類比層級轉換成數位位元)。

【0085】 一般而言，對於VMM陣列中之每一記憶體單元，每一權重 W 可由單個記憶體單元或差分單元或兩個混合記憶體單元(2個單元之平均值)實施。在差分單元情況下，需要兩個記憶體單元將權重 W 實施為差分權重($W = W^+ - W^-$)。在兩個混合記憶體單元中，需要兩個記憶體單元將權重 W 實施為兩個單元之平均值。

【0086】 類比神經形態記憶體系統中使用之每一非揮發性記憶體單元必須經抹除及程式化以在浮動閘極中保持極特定且精確的電荷量，亦即，電子數目。舉例而言，每一浮動閘極必須保持 N 個不同值中之一者，其中 N 為可由每一單元指示之不同權重之數目。 N 之實施例包括16、32、64、128及256。

【0087】 在VMM系統中需要儘可能地增加輸送量並減少潛時，同時減少記憶體單元及支援電路系統所需之總空間量。

【發明內容】

【0088】 揭示用於將一深度學習人工神經網路中之一類比神經記憶體中之一陣列分離成多個部分的眾多具體例，其中每一部分與專用於彼部分之某電路系統及與一或多個其他部分共用之其他電路系統互動。

【圖式簡單說明】

【0089】 圖1為例示人工神經網路之圖式。

【0090】 圖2描述先前技術分離閘極快閃記憶體單元。

【0091】 圖3描述另一先前技術分離閘極快閃記憶體單元。

【0092】 圖4描述另一先前技術分離閘極快閃記憶體單元。

【0093】 圖5描述另一先前技術分離閘極快閃記憶體單元。

【0094】 圖6為例示利用一或多個非揮發性記憶體陣列之例示性人工神經網路之不同層級的圖式。

【0095】 圖7為例示向量乘矩陣乘法系統之方塊圖。

【0096】 圖8為例示利用一或多個向量乘矩陣乘法系統之例示性人工神經網路的方塊圖。

【0097】 圖9描述向量乘矩陣乘法系統之另一具體例。

【0098】 圖10描述向量乘矩陣乘法系統之另一具體例。

【0099】 圖11描述向量乘矩陣乘法系統之另一具體例。

【0100】 圖12描述向量乘矩陣乘法系統之另一具體例。

【0101】 圖13描述向量乘矩陣乘法系統之另一具體例。

【0102】 圖14描述先前技術長短期記憶體系統。

【0103】 圖15描述用於長短期記憶體系統中之例示性單元。

【0104】 圖16描述圖15之例示性單元之具體例。

【0105】 圖17描述圖15之例示性單元之另一具體例。

【0106】 圖18描述先前技術閘控遞回單元系統。

【0107】 圖19描述用於閘控遞回單元系統中之例示性單元。

【0108】 圖20描述圖19之例示性單元之具體例。

【0109】 圖21描述圖19之例示性單元之另一具體例。

【0110】 圖22描述向量乘矩陣乘法系統之另一具體例。

【0111】 圖23描述向量乘矩陣乘法系統之另一具體例。

【0112】 圖24描述向量乘矩陣乘法系統之另一具體例。

【0113】 圖25描述向量乘矩陣乘法系統之另一具體例。

【0114】 圖26A描述向量乘矩陣乘法系統之另一具體例。

【0115】 圖26B描述向量乘矩陣乘法系統之另一具體例。

【0116】 圖26C描述向量乘矩陣乘法系統之另一具體例。

【0117】 圖27描述向量乘矩陣乘法系統之另一具體例。

【0118】 圖28描述向量乘矩陣乘法系統之另一具體例。

【0119】 圖29描述向量乘矩陣乘法系統之另一具體例。

【0120】 圖30描述向量乘矩陣乘法系統之另一具體例。

【0121】 圖31描述向量乘矩陣乘法系統。

【0122】 圖32描述分離之向量乘矩陣乘法系統之具體例。

【0123】 圖33描述分離之陣列向量乘矩陣乘法系統之具體例。

【0124】 圖34描述分離陣列向量乘矩陣乘法系統之另一具體例。

【0125】 圖35描述分離陣列向量乘矩陣乘法系統之另一具體例。

【0126】 圖36描述分離陣列向量乘矩陣乘法系統之另一具體例。

【0127】 圖37描述向量乘矩陣乘法系統中之分離陣列之具體例。

【0128】 圖38描述向量乘矩陣乘法系統中之分離陣列之另一具體例。

【0129】 圖39描述向量乘矩陣乘法系統中之單個陣列及分離陣列之例示性佈局。

【實施方式】

【0130】 本發明之人工神經網路利用CMOS技術及非揮發性記憶體陣列之組合。

VMM系統綜述

【0131】 圖31描述VMM系統3100之方塊圖。VMM系統3100包含VMM陣列3101、列解碼器3102、高電壓解碼器3103、行解碼器3104、位元線驅動器3105、輸入電路3106、輸出電路3107、控制邏輯3108及偏壓產生器3109。VMM系統3100進一步包含高電壓產生區塊3110，該高電壓產生區塊包含電荷泵3111、電荷泵調節器3112及高電壓位準產生器3113。VMM系統3100進一步包含(程式化/抹除，或亦稱為權重調諧)演算法控制器3114、類比電路系統3115、控制引擎3116 (其可包括特殊函數，諸如算術函數、激勵函數、嵌入式微控制器邏輯等)，及測試控制邏輯3117。下文描繪之系統及方法可實施於VMM系統3100中。

【0132】 輸入電路3106可包括電路，諸如數位至類比轉換器(digital to analog converter；DAC)、數位至脈衝轉換器(digital to pulses converter；DPC、數位至時間經調變脈衝轉換器)、類比至類比轉換器(analog to analog converter；AAC，諸如電流至電壓轉換器、對數轉換器)、脈衝至類比層級轉換器(pulse to analog level converter；PAC)，或任何其他類型之轉換器。輸入電路3106可實施正規化、線性或非線性按比例放大/按比例縮小函數，或算術函數。輸入電路3106可實施用於輸入層級之溫度補償函數。輸入電路3106可實施激勵函數，諸如ReLU或S型。輸出電路3107可包括電路，諸如類比至數位轉換器(analog to digital converter；ADC，其用以將神經元類比輸出轉換成數位位元)、

類比至類比轉換器(analog to analog converter；AAC，諸如電流至電壓轉換器、對數轉換器)、類比至脈衝轉換器(analog to pulse(s) converter；APC，類比至時間經調變脈衝轉換器)，電流至電壓轉換器或任何其他類型之轉換器。輸出電路3107可實施諸如ReLU或S型之激勵函數。輸出電路3107可實施統計正規化、正則化、按比例放大/按比例縮小/增益函數，統計捨位或算術函數(例如，加法、減法、除法、乘法、移位、對數)以用於神經元輸出。輸出電路3107可實施溫度補償功能以用於神經元輸出或陣列輸出(諸如位元線輸出)，以便使陣列之功耗保持近似恆定或諸如藉由使IV斜率保持大致相同而改良陣列(神經元)輸出之精度。

【0133】 圖32至圖36描述VMM系統之具體例，VMM系統與VMM系統3100含有一定的共同性但亦含有一些修改。

【0134】 圖32描述VMM系統3200。VMM系統3200包含陣列3201、共用列解碼器3202、共用高電壓解碼器3203、行解碼器3204及3205、(列)輸入電路3220、輸出電路3206及3207以及共用位元線驅動器3208。共用列解碼器3202耦接至陣列3201中之所有列且將電壓施加至選定列。共用高電壓解碼器3203可選擇性地耦接至陣列3201中之所有列。共用高電壓解碼器3203選擇地包含可選擇性地耦接至陣列中之所有列之控制閘極高電壓解碼器3231，及可選擇性地耦接至陣列中之所有列之共用抹除閘極高電壓解碼器3232。輸入電路3220例如類似於圖31之輸入電路3106。輸出電路3206及3207之電路及功能例如各自類似於圖31之輸出電路3107的電路及功能。不同於VMM系統3100中，在VMM系統3200中，某些操作在不同電路系統集合之間分離。特定言之，陣列3201中之一半

的行(例如所有奇數行)由行解碼器3204及輸出電路3206操作，且陣列3201中之另一半行(例如所有偶數行)係由行解碼器3205及輸出電路3207操作。因此，輸出電路3206耦接至行解碼器3204以用於在讀取操作期間自行之前半部分中的一或多個行產生第一輸出，且輸出電路3207耦接至行解碼器3205以用於在讀取操作期間自行之後半部分中的一或多個行產生第二輸出。在此具體例中，所有行在程式化或抹除操作期間耦接至共用位元線驅動器3208。此允許同時讀取多個位元線，意謂耦接至行解碼器3204及輸出電路3206之位元線及耦接至行解碼器3205及輸出電路3207之位元線藉由共用位元線驅動器3208同時啟用，以用於讀取操作。因此，此增加了讀取陣列3201之輸送量。替代地，讀取操作無需為同時的。

【0135】 選擇地，進一步參見圖39，連續擴散可實施於陣列之頂半部分與底半部分之間。

【0136】 圖33描述VMM系統3300。VMM系統3300包含陣列3301a及3301b、列解碼器3302、共用高電壓解碼器3303、行解碼器3304及3305、輸入電路3320、電流至電壓轉換器電路3306及3307、共用類比至數位轉換器(ADC) 3308及共用位元線驅動器3309。電流至電壓轉換器電路3306或3307及共用ADC電路3308為圖32中之輸出電路3207之部分。

【0137】 不同於VMM系統3100中，在VMM系統3300中，某些操作在不同電路系統集合之間分離。特定言之，陣列3301a係由行解碼器3304及電流至電壓轉換器3306操作，且陣列3301b係由行解碼器3305及電流

至電壓轉換器3307操作。此允許同時執行多個讀取及/或程式化操作，其中可對陣列3301a中之一或多個單元及陣列3301b中之一或多個單元同時執行讀取或程式化操作。

【0138】 電流至電壓轉換器電路3306及3307兩者耦接至共用類比至數位轉換器3308，其在讀取操作期間以時間多工方式使用，且耦接至共用位元線驅動器3309，其在程式化及抹除操作期間使用。舉例而言，在讀取操作中，陣列3301a經啟用且耦接至行解碼器3304及電流至電壓轉換器電路3306，同時陣列3301b經啟用且同時耦接至行解碼器3305及電流至電壓轉換器電路3307。來自電流至電壓轉換器電路3306及3307之輸出電壓例如藉由共用ADC 3308內部之S/H電容器而取樣及保持(S/H)，且此等陣列輸出電壓藉由時間多工共用之ADC 3308而數位化(轉換)(此係由於該ADC在電流至電壓轉換器電路3306與3307之間共用)。舉例而言，對於兩個電流至電壓轉換器電路之間共用的一個ADC，使用兩組S/H電容器。在另一具體例中，一個ADC可用於N個電流至電壓轉換器電路，且在此情況下，使用N組S/H電容器。

【0139】 在兩個電流至電壓轉換器電路之間使用共用ADC亦可適用於圖34/圖35/圖36。

【0140】 圖34描述VMM系統3400。VMM系統3400包含陣列3401a及3401b、共用列解碼器3402、共用高電壓解碼器3403、行解碼器3404及3405、輸入電路3420、輸出電路3406及3407以及共用位元線驅動器3408。不同於VMM系統3100中，在VMM系統3400中，某些操作在不同電路系統集合之間分離。特定言之，陣列3401a係由行解碼器3404及輸

出電路3406操作，且陣列3401b係由行解碼器3405及輸出電路3407操作。此允許同時執行多個讀取/或及程式化操作，其中可對陣列3401a中之一或多個單元及陣列3401b中之一或多個單元同時執行讀取或程式化操作。陣列3401a及3401b兩者耦接至共用位元線驅動器3408，共用位元線驅動器在程式化及抹除操作期間使用。

【0141】 圖35描述VMM系統3500。VMM系統3500包含陣列3501a、3501b、3501c及3501d；列解碼器3502及3503；共用高電壓解碼器3504；行解碼器3505、3506、3507及3508；輸入電路3520；輸出電路3509、3510、3511及3512；以及共用位元線驅動器3513及3514。共用高電壓解碼器3504可選擇性地耦接至陣列3501a、3501b、3501c及3501d中之所有列。列解碼器3502由陣列3501a及3501b共用且耦接至彼等陣列中之所有列且將電壓施加至選定列，且列解碼器3503由陣列3501c及3501d共用且耦接至彼等陣列中之所有列且將電壓施加至選定列。

【0142】 在VMM系統3500中，某些操作在不同電路系統集合之間分離。特定言之，陣列3501a係由行解碼器3505及輸出電路3509操作；陣列3501b係由行解碼器3507及輸出電路3511操作；陣列3501c係由行解碼器3506及輸出電路3510操作；且陣列3501d係由行解碼器3508及輸出電路3512操作。此允許同時在所有四個陣列中同時執行多個讀取/或及程式化操作，其中可同時對陣列3501a中之一或多個單元、陣列3501b中之一或多個單元、陣列3501c中之一或多個單元及陣列3501d中之一或多個單元同時執行讀取或程式化操作。陣列3501a及3501b在程式化及抹除

操作期間均選擇性地耦接至共用位元線驅動器3513。陣列3501c及3501d兩者在程式化及抹除操作期間選擇性地耦接至共用位元線驅動器3514。

【0143】舉例而言，可執行第一讀取操作，其中行解碼器3505及輸出電路3509自陣列3501a中之一或多個列產生第一輸出；可執行第二讀取操作，其中行解碼器3506及輸出電路3510自陣列3501c中之一或多個列產生第二輸出；可執行第三讀取操作，其中行解碼器3507及輸出電路3511自陣列3501b中之一或多個列產生第三輸出；且可執行第四讀取操作，其中行解碼器3508及輸出電路3512自陣列3501d中之一或多個列產生第四輸出。選擇地，第一及第三讀取操作可同時發生。選擇地，第二及第四讀取操作可同時發生。

【0144】圖36描述VMM系統3600。VMM系統3600包含陣列3601a、3601b、3601c及3601d；列解碼器3621；控制閘極解碼器3602及3603；共用高電壓解碼器3604；行解碼器3605、3606、3607及3608；輸出電路3609、3610、3611及3612；以及共用位元線驅動器3613及3614。在VMM系統3600中，某些操作在不同電路系統集合之間分離。特定言之，陣列3601a係由行解碼器3605及輸出電路3609操作；陣列3601b係由行解碼器3607及輸出電路3611操作；陣列3601c係由行解碼器3606及輸出電路3610操作；且陣列3601d係由行解碼器3608及輸出電路3612操作。此允許同時在所有四個陣列中同時執行多個讀取及/或程式化操作，其中可同時對陣列3601a中之一或多個單元、陣列3601b中之一或多個單元、陣列3601c中之一或多個單元及陣列3601d中之一或多個單元同時執行讀取或程式化操作。陣列3601a及3601b兩者在程式化及抹除操作期間選

擇性地耦接至共用位元線驅動器3613。陣列3601c及3601d兩者在程式化及抹除操作期間選擇性地耦接至共用位元線驅動器3614。

【0145】 圖32至圖36顯示讀取係藉由控制閘極上之列輸入來完成。替代地，讀取可在字線或抹除閘極上完成。圖32中之輸入電路3220、圖33中之輸入電路3320、圖34中之輸入電路3420、圖35中之輸入電路3520及圖36中之輸入電路3620類似於圖31之輸入電路3106。圖32中之輸出電路3206/3207及圖34中之輸出電路3406/3407、圖35中之輸出電路3507/3508/3509/3510以及圖36中之輸出電路3607/3608/3609/3610類似於圖31之輸出電路3107。

【0146】 圖37描述VMM陣列3700之一部分。VMM陣列3700包含列3701、3702、3703、3704、3705、3706、3707及3708。列3701、3702、3705及3706共用抹除閘極線(EG0)及源極線(SL0)；列3703、3704、3707及3708共用抹除閘極線(EG1)及源極線(SL1)。另外，列3701及3703共用控制閘極線(CG0/CG2)；列3702及3704共用控制閘極線(CG1/CG3)；列3705及3707共用控制閘極線(CG4/CG6)；且列3706及3708共用控制閘極線(CG5/CG7)。此等耦接允許不同列共用解碼器電路系統。陣列端子經共用使得藉由在未選定單元上具有減少量之抹除或程式化電壓應力來減少程式化或抹除干擾。

【0147】 在圖37及圖38(下文所描繪)之陣列中，用於神經讀取操作之VMM陣列3700及3800之列輸入在字線上開啟(多個列及多個位元線同時開啟)。若用於神經讀取之輸入在控制閘極上開啟，則控制閘極無法跨越同一子陣列或陣列組中之多個列被共用。

【0148】 圖38描述陣列3800之一部分。陣列3800包含扇區3809及3819。扇區3809包含列3801、3802、3803、3804、3805、3806、3807及3808。扇區3819包含列3811、3812、3813、3814、3815、3816、3817及3818。

【0149】 列3801(第一列)及3811(第二列)共用控制閘極線(CG0) (意謂彼等列中之每一單元的控制閘極端子耦接至同一控制閘極線)；列3802及3812共用控制閘極線(CG1) (意謂彼等列中之每一單元的控制閘極端子耦接至同一控制閘極線)；列3803及3813共用控制閘極線(CG2) (意謂彼等列中之每一單元的控制閘極端子耦接至同一控制閘極線)；列3804及3814共用控制閘極線(CG3) (意謂彼等列中之每一單元的控制閘極端子耦接至同一控制閘極線)；列3805及3815共用控制閘極線(CG4) (意謂彼等列中之每一單元的控制閘極端子耦接至同一控制閘極線)；列3806及3816共用控制閘極線(CG5) (意謂彼等列中之每一單元的控制閘極端子耦接至同一控制閘極線)；列3807及3817共用控制閘極線(CG6) (意謂彼等列中之每一單元的控制閘極端子耦接至同一控制閘極線)；並且列3808及行3818共用控制閘極線(CG7) (意謂彼等列中之每一單元的控制閘極端子耦接至同一控制閘極線)。此意謂控制閘極跨越扇區共用。此等耦接允許不同列共用解碼器電路系統。陣列端子經共用使得藉由在未選定單元上具有減少量之抹除或程式化電壓應力來減少程式化或抹除干擾。

【0150】 列3801 (第一列)、3802 (第三列)、3805及3806共用抹除閘極線(EG0) (意謂彼等列中之每一單元的抹除閘極端子耦接至同一抹

除閘極線)及源極線(SL0) (意謂彼等列中之每一單元的源極線端子耦接至同一源極線);列3803、3804、3807及3808共用抹除閘極線(EG1) (意謂彼等列中之每一單元的抹除閘極端子耦接至同一抹除閘極線)及源極線(SL1) (意謂彼等列中之每一單元的源極線端子耦接至同一源極線);列3811、3812、3815及3816共用抹除閘極線(EG0) (意謂彼等列中之每一單元的抹除閘極端子耦接至同一抹除閘極線)及源極線(SL0) (意謂彼等列中之每一單元的源極線端子耦接至同一源極線);並且列3813、3814、3817及3818共用抹除閘極線(EG1) (意謂彼等列中之每一單元的抹除閘極端子耦接至同一抹除閘極線)及源極線(SL1) (意謂彼等列中之每一單元的源極線端子耦接至同一源極線)。

【0151】 圖39描述單個陣列3901 (諸如圖31中之陣列3101及圖32中之陣列3201)及分離陣列3902 (諸如圖33中之陣列3301a及3301b、圖34中之陣列3401a及3401b、圖35中之陣列3501a、3501b、3501c及3501d以及圖36中之陣列3601a、3601b、3601c及3601d)之一部分的例示性佈局。分離陣列3902遵循與陣列3901相同之設計，其例外之處在於某些接點及金屬連接3904被移除(或未形成)，從而創建子陣列3903a及3903b。介面處之少數虛設列諸如藉由使字線及控制閘極接地而停用。此由於前端層(亦即，行內之連續行擴散及源極線內之連續列擴散)及多晶矽在非揮發性記憶體單元之兩個陣列之間(電分離陣列之間)係連續及均一的從而維持程序均一性。此亦導致與不同陣列之實體分離相比，面積開銷減小。

【0152】 應注意，如本文中所使用，術語「在…上方」及「在…上」兩者包括「在…正上方」(其間未裝設有中間材料、元件或空間)及「間接地在…上」(其間裝設有中間材料、元件或空間)。同樣地，術語「鄰近」包括「直接鄰近」(其間未裝設有中間材料、元件或空間)及「間接鄰近」(其間裝設有中間材料、元件或空間)，「安裝至」包括「直接安裝至」(其間未裝設有中間材料、元件或空間)及「間接安裝至」(其間裝設有中間材料、元件或空間)，且「電耦接」包括「直接電耦接至」(其間不具有將元件電連接在一起的中間材料或元件)及「間接電耦接至」(其間具有將元件電連接在一起的中間材料或元件)。舉例而言，「在基板上方」形成元件可包括在基板正上方形成元件而其間不具有中間材料/元件，以及間接地在基板上形成元件，其間具有一或多種中間材料/元件。

【符號說明】

【0153】

12:半導體基板

14:源極區

16:汲極區

18:通道區

20:浮動閘極

22:字線端子/選擇閘極

28:控制閘極

30:抹除閘極

31:數位至類比轉換器

- 32:輸入向量乘矩陣乘法(VMM)陣列
- 32a:向量乘矩陣乘法(VMM)陣列/輸入層
- 32b,32c:向量乘矩陣乘法(VMM)陣列/隱藏層
- 32d,32e:向量乘矩陣乘法(VMM)陣列/完全連接層
- 33:非揮發性記憶體單元陣列
- 34:抹除閘極及字線閘極解碼器
- 35,3602,3603:控制閘極解碼器
- 36:位元線解碼器
- 37:源極線解碼器
- 38:差分求和器
- 39,1602,1702,2002,2102:激勵函數區塊
- 210:記憶體單元/快閃記憶體單元
- 310:四閘極記憶體單元/快閃記憶體單元
- 410:三閘極記憶體單元/快閃記憶體單元
- 510:堆疊閘極記憶體單元/快閃記憶體單元
- 900,1000,1100,1200,1300,2200,2300,2400,2500,2600,2620,
2700,2800,2900,3000:神經元輸入向量乘矩陣乘法(VMM)陣列
- 901,1003,1103,1203,1303:非揮發性記憶體單元之記憶體陣列
- 902:非揮發性參考記憶體單元之參考陣列
- 903:控制閘極線
- 904:抹除閘極線
- 1001,1101,1201,1301:第一非揮發性參考記憶體單元之參考陣列

1002,1102,1202,1302:第二非揮發性參考記憶體單元之參考陣列

1014,1212:二極體連接式貫穿多工器

1204:串疊電晶體

1205,1709,1710,2104,2106,2107,2108:多工器

1314:二極體連接式參考單元貫穿多工器

1400:長短期記憶體(LSTM)

1401,1402,1403,1404,1801,1802,1803,1804:單元

1500,1600,1700:長短期記憶體(LSTM)單元

1501,1502,1503,1901,1902:S型函數構件

1504,1505,1903:雙曲正切構件

1506,1507,1508,1703,1904,1905,1906,2103:乘法器構件

1509,1708,1907,2105:加法構件

1601,1701,2001,2101,3101,3700:向量乘矩陣乘法(VMM)陣列

1704,1705,1706,1707:暫存器

1800:閘控遞回單元(GRU)

1900,2000,2100:閘控遞回單元(GRU)單元

1908,2109:互補構件

2701-1,2701-2,2701-(N-1),2701-N:位元線控制閘極

3100,3200,3300,3400,3500,3600:向量乘矩陣乘法(VMM)系統

3102,3302,3502,3503,3621:列解碼器

3103:高電壓解碼器

3104,3204,3205,3304,3305,3404,3405,3505,3506,3507,3508,

3605,3606,3607,3608:行解碼器

3105:位元線驅動器

3106,3220,3320,3420,3520,3620:輸入電路

3107,3206,3207,3406,3407,3509,3510,3511,3512,3609,3610,
3611,3612:輸出電路

3108:控制邏輯

3109:偏壓產生器

3110:高電壓產生區塊

3111:電荷泵

3112:電荷泵調節器

3113:高電壓位準產生器

3114:演算法控制器

3115:類比電路系統

3116:控制引擎

3117:測試控制邏輯

3201,3301a,3301b,3401a,3401b, 501a,3501b,3501c,3501d,3800:
陣列

3202,3402:共用列解碼器

3203,3303,3403,3504,3604:共用高電壓解碼器

3208,3309,3408,3513,3514,3613,3614:共用位元線驅動器

3231:控制閘極高電壓解碼器

3232:共用抹除閘極高電壓解碼器

3306,3307: 電流至電壓轉換器電路

3308: 共用類比至數位轉換器(ADC)

3601a,3601b,3601c,3601d: 陣列

3701,3702,3703,3704,3705,3706,3707,3708,3801,3802,3803,

3804,3805,3806,3807,3808,3811,3812,3813,3814,3815,3816,3817,

3818: 列

3809,3819: 扇區

3901: 單個陣列

3902: 分離陣列

3903a,3903b: 子陣列

3904: 接點及金屬連接

BL0,BL1,BL2,BL3,BLN,24: 位元線

BLx,CGx,EGx,Inputx,INPUT₀,INPUT₁,INPUT_{N-1},INPUT_N,WLx:
輸入

BLR0,BLR1,BLR2,BLR3: 端子

C1,C2,C3,S1,S2: 層

CB1,CB2,CB3,CB4: 突觸

CG0,CG1,CG2,CG3,CG4,CG5,CG6,CG7,CG_{M-1},CG_M: 控制閘極線
/電壓輸入/控制閘極

c₀,c₁,c₂,c(t-1),c(t): 單元狀態向量

EG0,EG1: 抹除閘極/抹除閘極(EG)線

EGR0,EGR1:抹除閘極(EG)線

$h_0,h_1,h_2,h_3,h(t-1),h(t)$:輸出向量

OUTPUT₁,OUTPUT₂,OUTPUT₃,OUTPUT₄:輸出

P1,P2:激勵函數/池化函數

S0:輸入層

S3:輸出層

SL0,SL1,SL2,SL3:源極線

WL,WLA0,WLA1,WLA2,WLA3,WLB0,WLB1,WLB2,WLB3:字

線

WL0,WL1,WL2,WL3,WL_{M-1},WL_M:字線/SL線

WLR0,WLR1,WLR2,WLR3:電壓輸入

$x_0,x_1,x_2,x_3,x(t)$:輸入向量

【發明申請專利範圍】

【請求項1】 一種類比神經記憶體，其包含：

一陣列的非揮發性記憶體單元，其配置成多列及多行，每個非揮發性記憶體單元包含一位元線端子、一控制閘極端子及一抹除閘極端子；

一第一行解碼器，其耦接至該陣列中之偶數行；

一第二行解碼器，其耦接至該陣列中之奇數行；

一共用位元線驅動器，其用以同時啟用經由個別非揮發性記憶體單元之位元線端子耦接至該第一行解碼器之多個行及經由個別非揮發性記憶體單元之位元線端子耦接至該第二行解碼器之多個行；

一共用控制閘極高電壓解碼器(3231)，其選擇性地耦接至該陣列中之多個列的控制閘極端子；

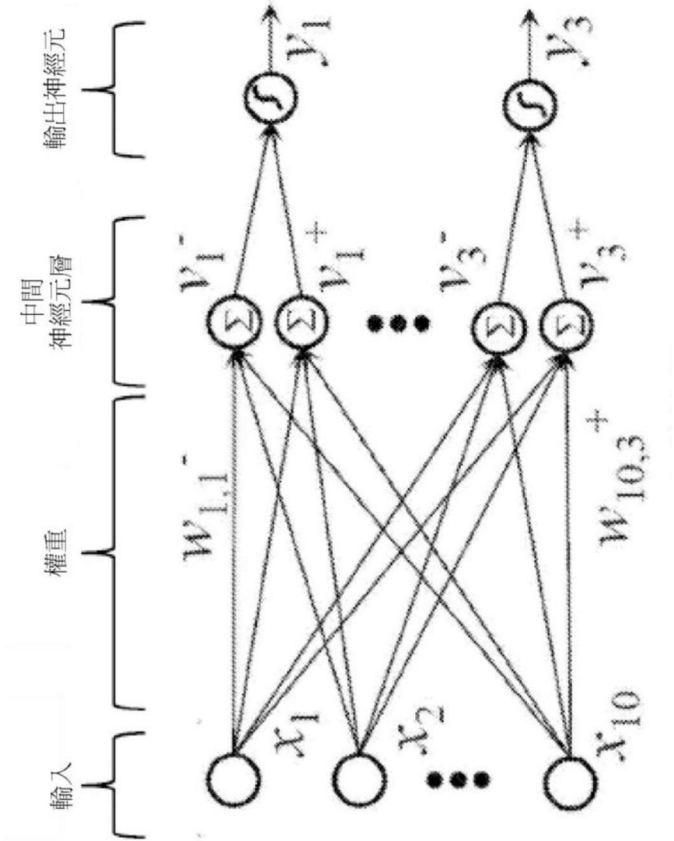
一共用抹除閘極高電壓解碼器(3232)，其選擇性地耦接至該陣列中之多個列的抹除閘極端子；

一共用列解碼器(3202)，其耦接至該陣列中之所有列；

一第一輸出電路，其耦接至該第一行解碼器以用於回應一或多個該等偶數行的該共用位元線驅動器所進行之啟用而產生一第一輸出；及

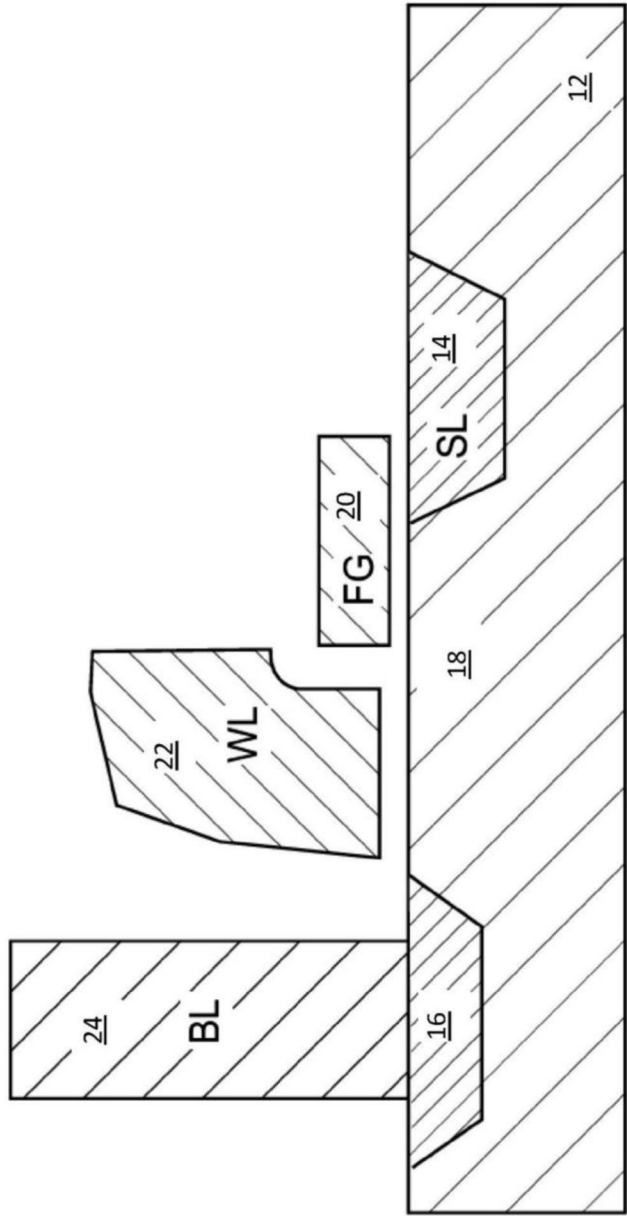
一第二輸出電路，其耦接至該第二行解碼器以用於回應一或多個該等奇數行的該共用位元線驅動器所進行之啟用而產生一第二輸出，其中該第一輸出和該第二輸出係同時產生。

【發明圖式】

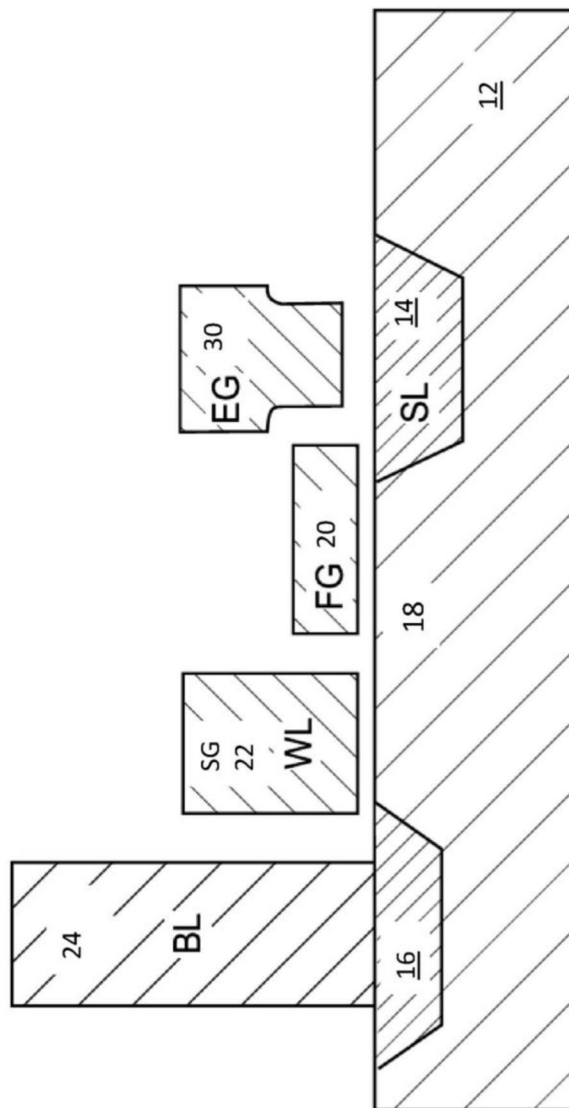


【圖1】

記憶體單元
210

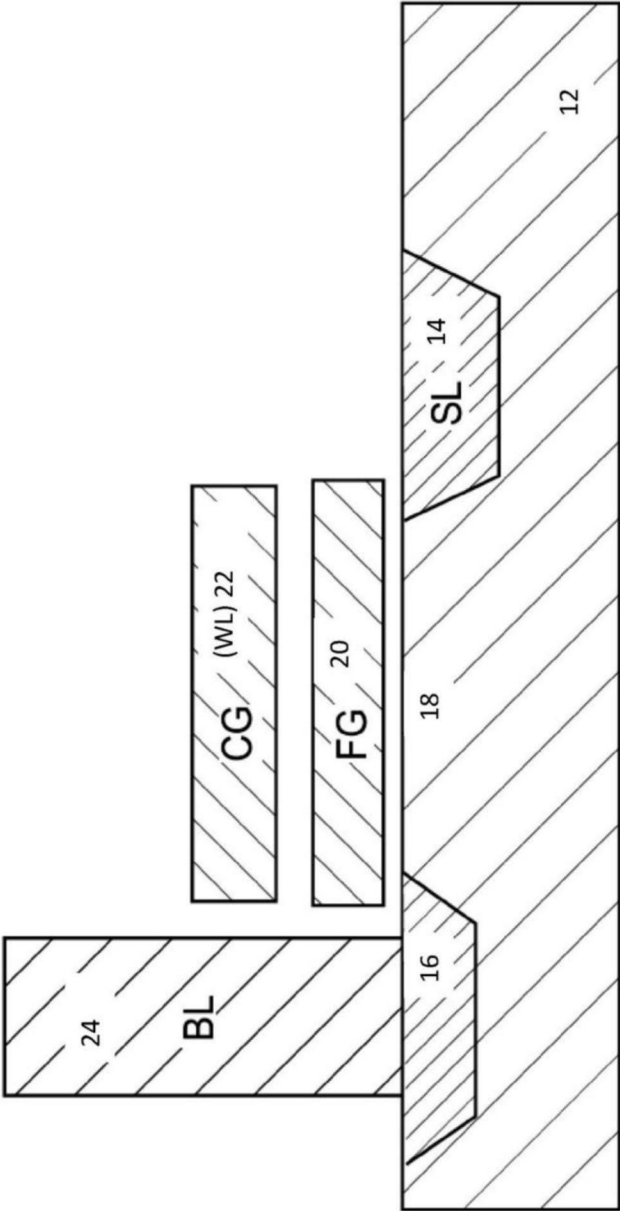


【圖2】

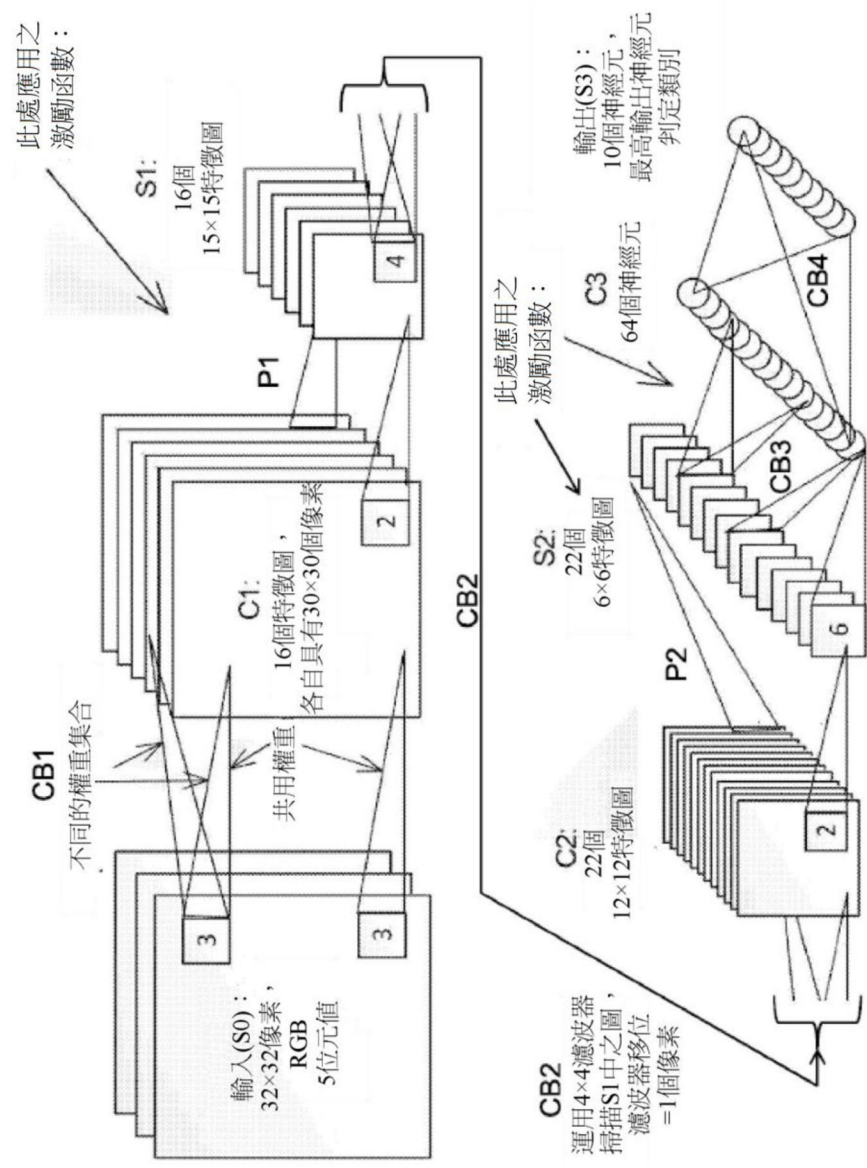


【4回】

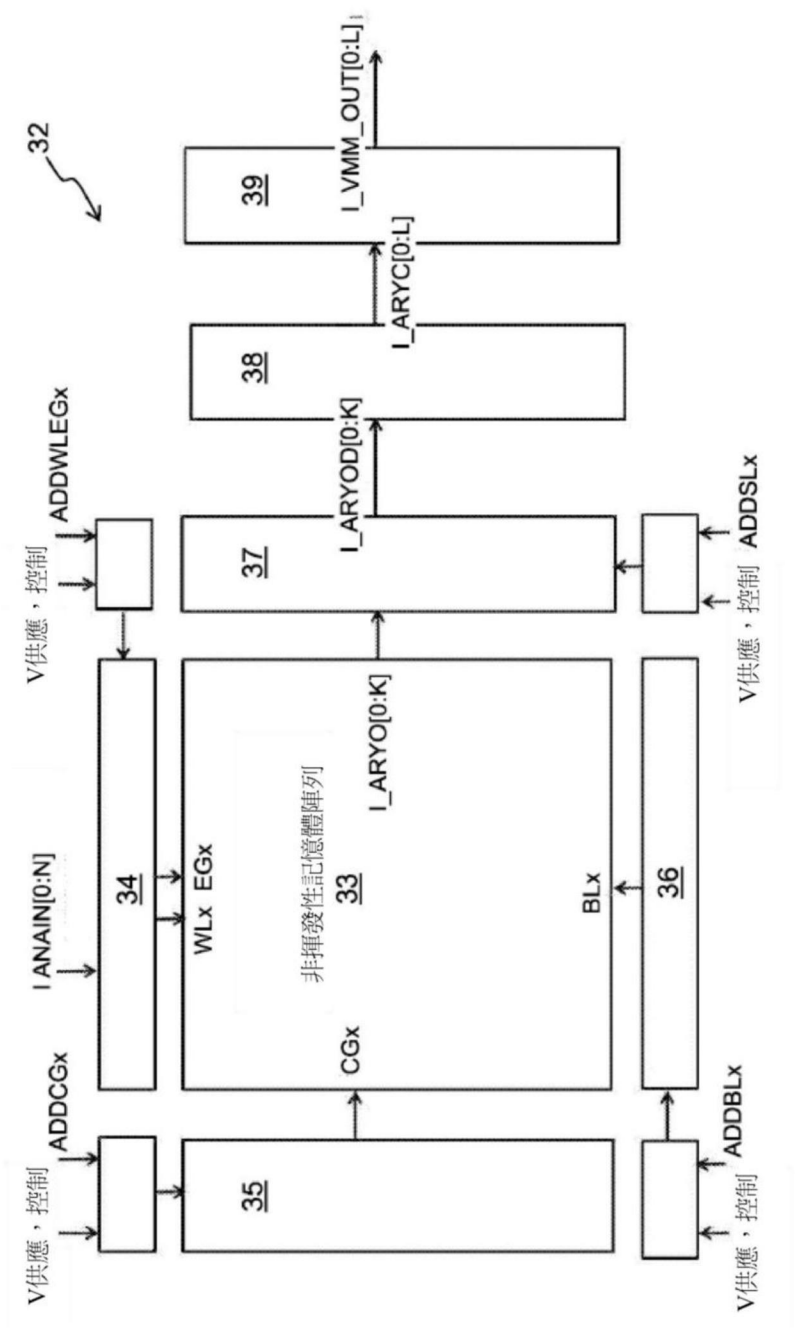
記憶體單元
510



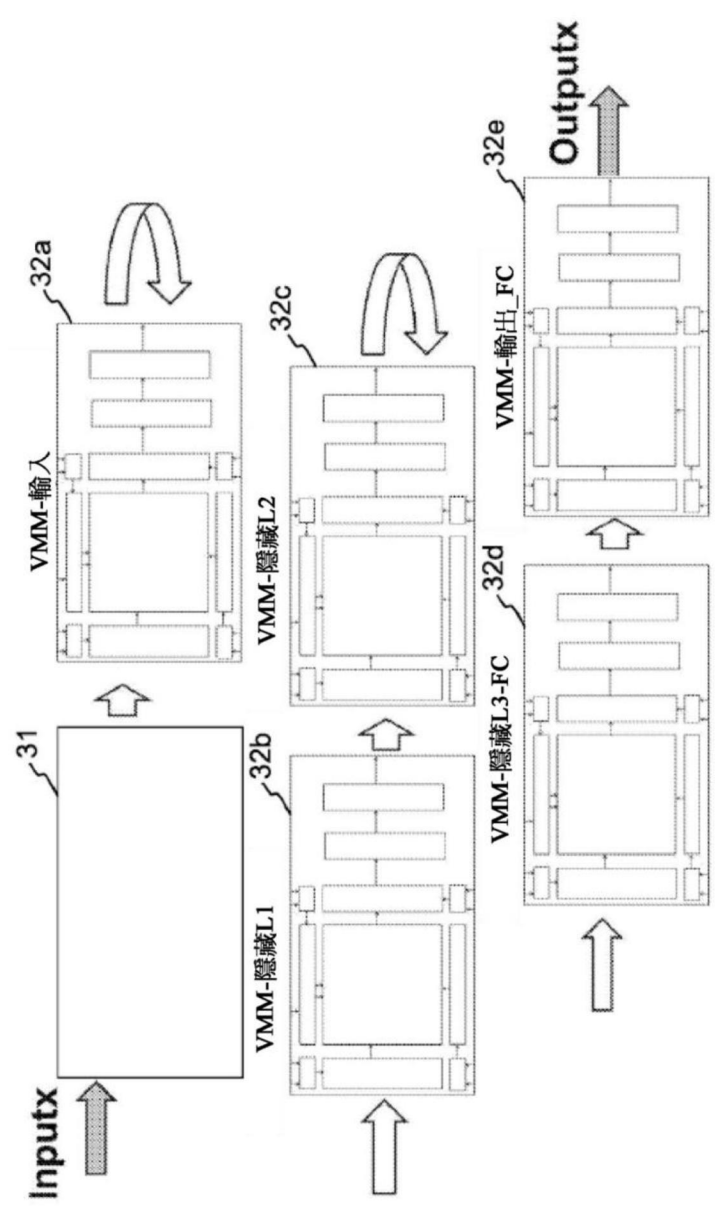
【圖5】



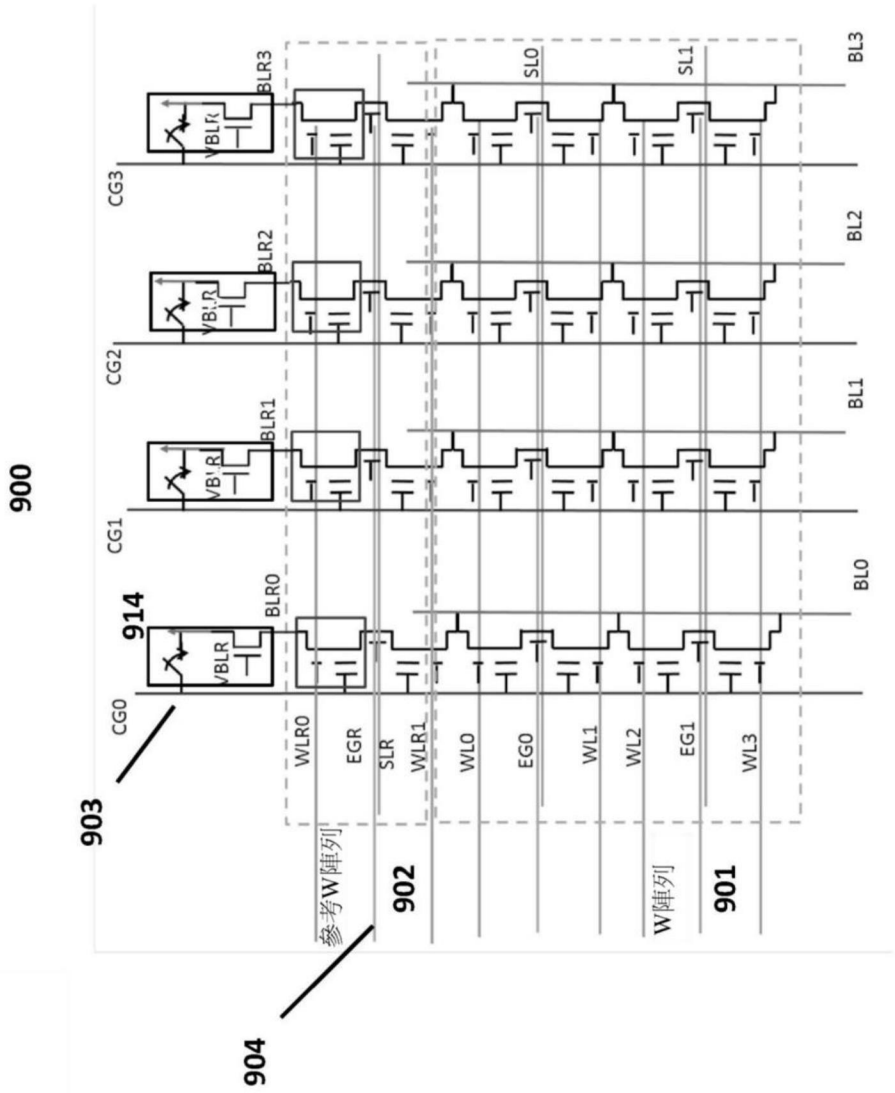
【圖6】



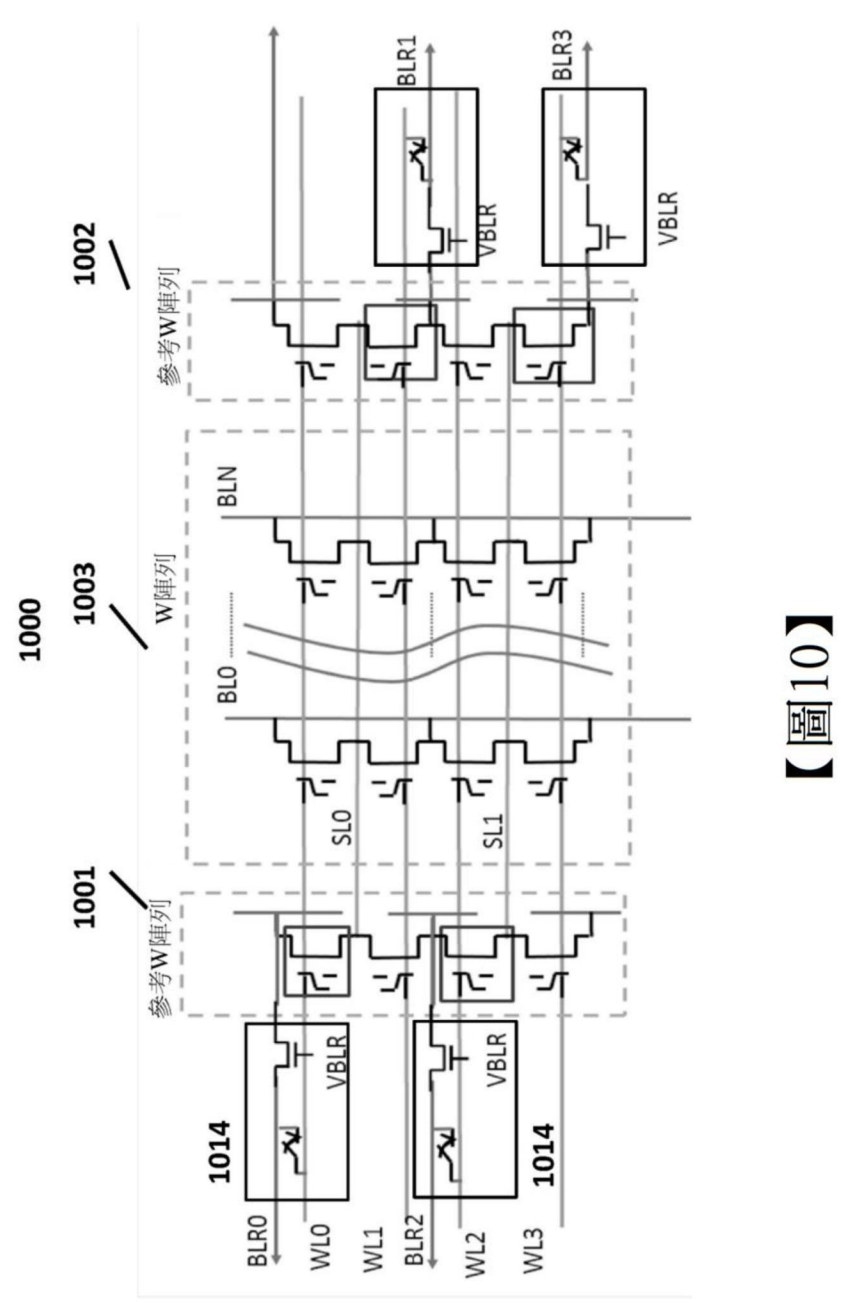
【圖7】



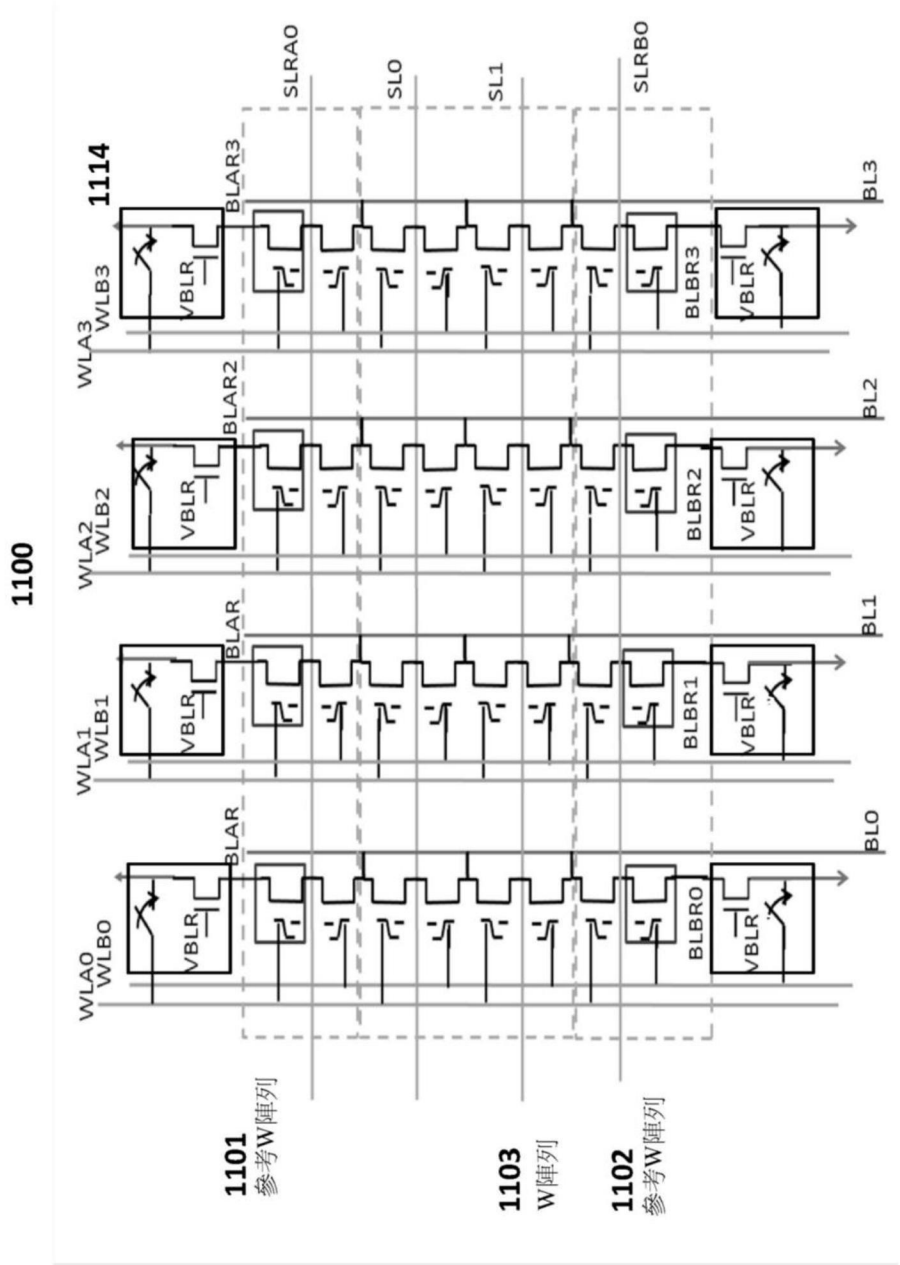
【圖8】



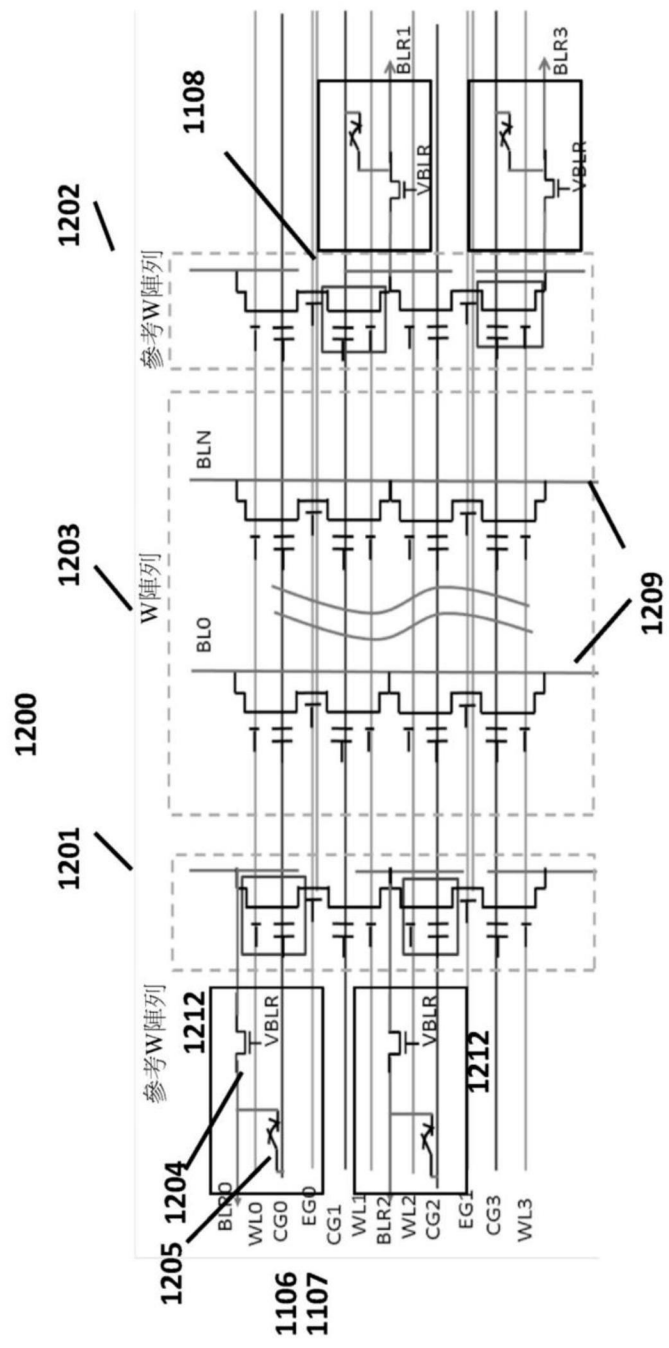
【圖9】



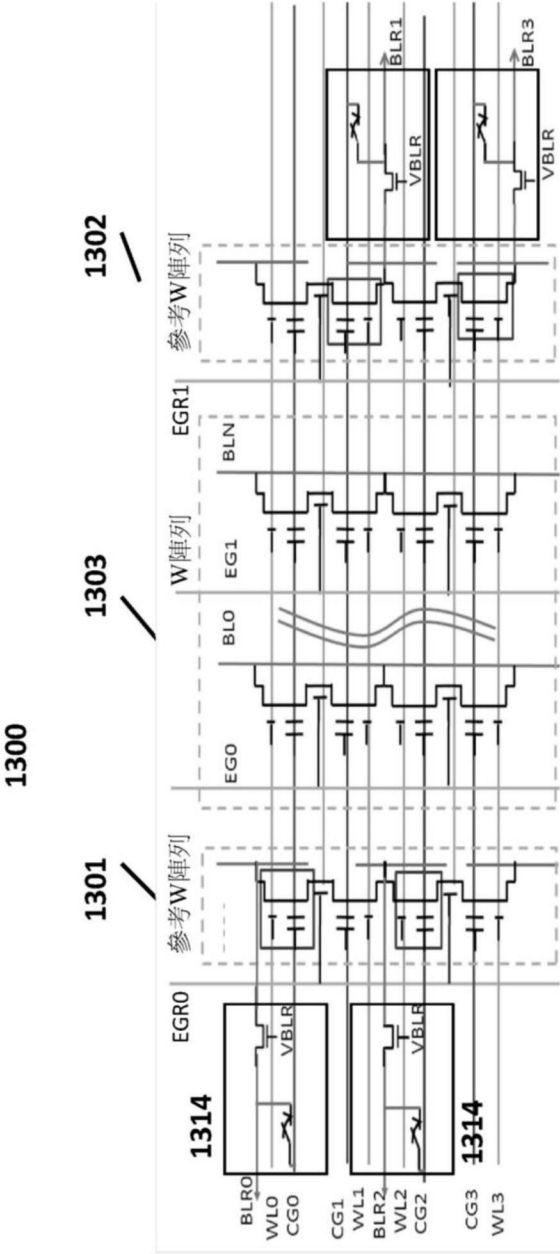
【圖10】



【圖11】

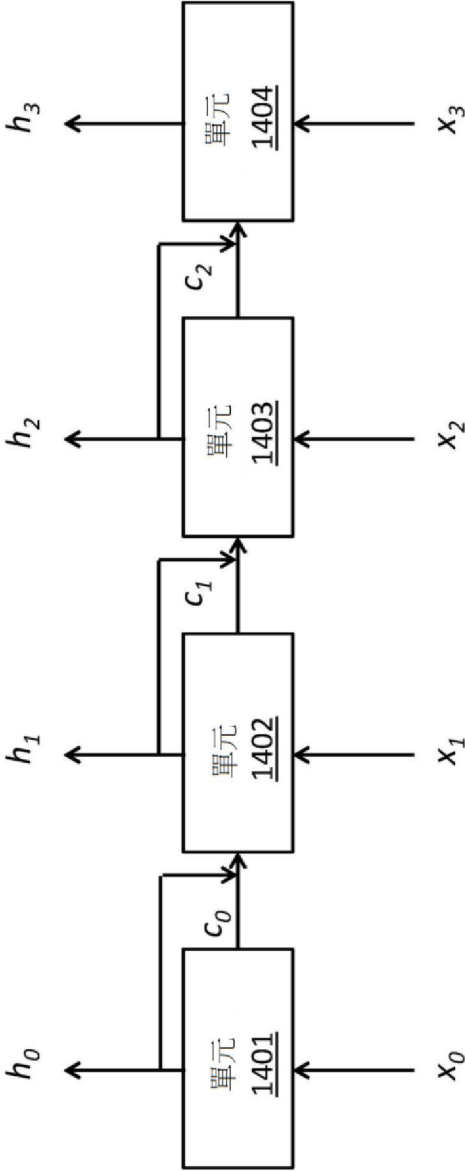


【圖12】

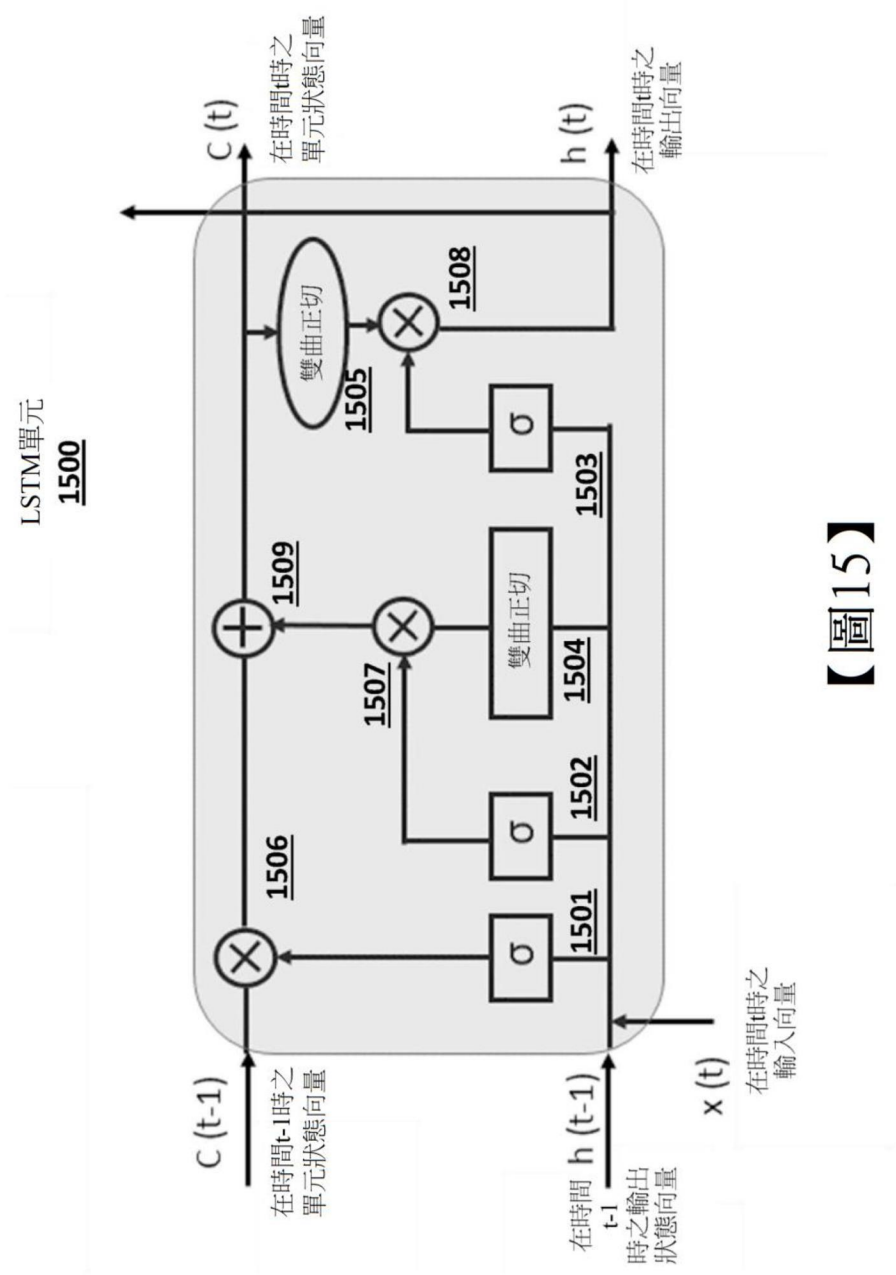


【圖13】

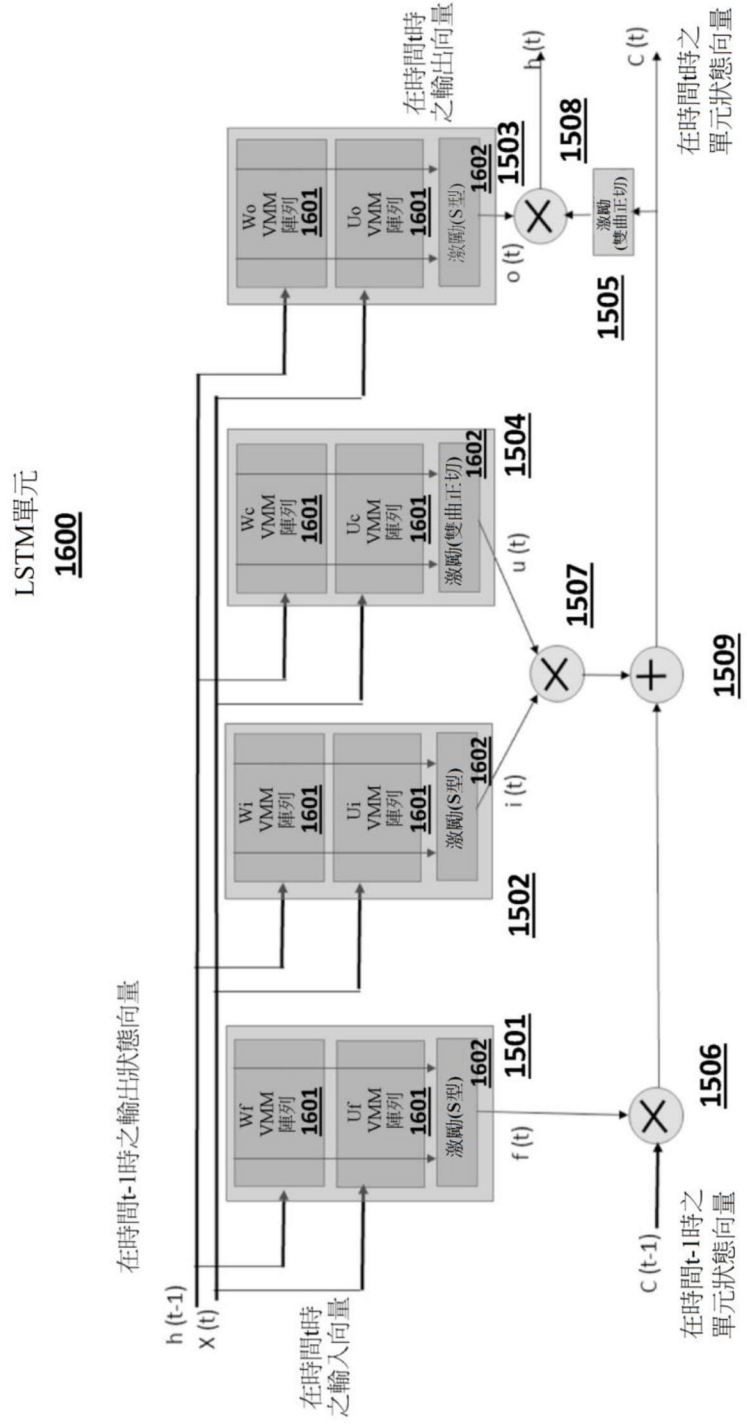
LSTM
1400



【圖14】

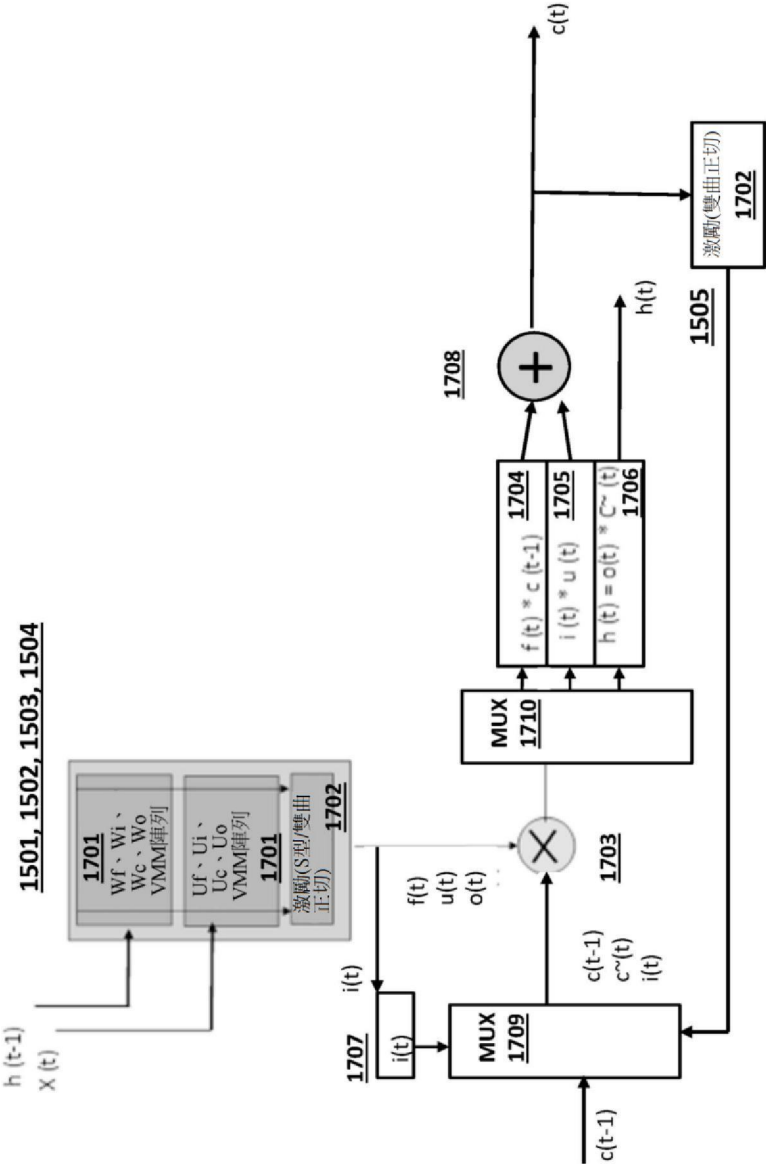


【圖15】



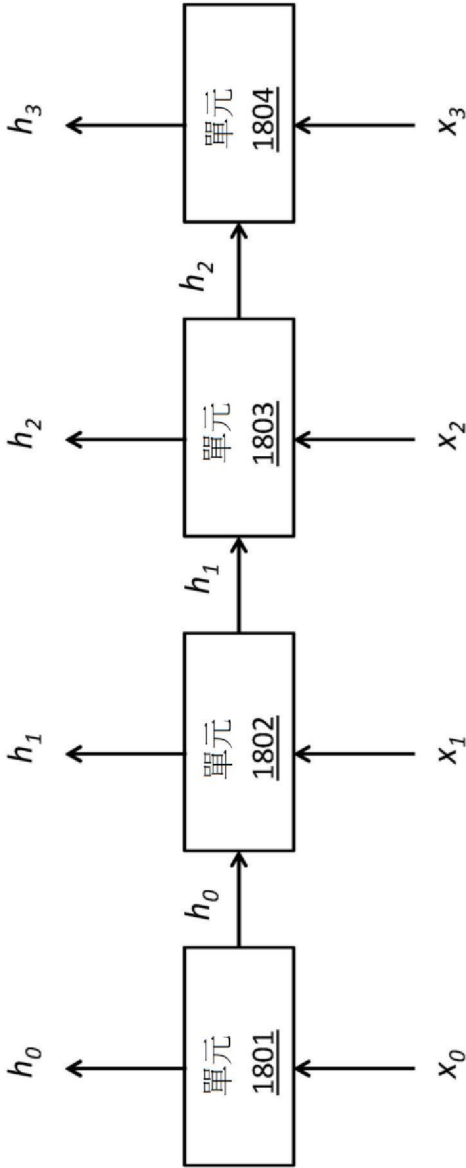
【圖16】

LSTM單元
1700

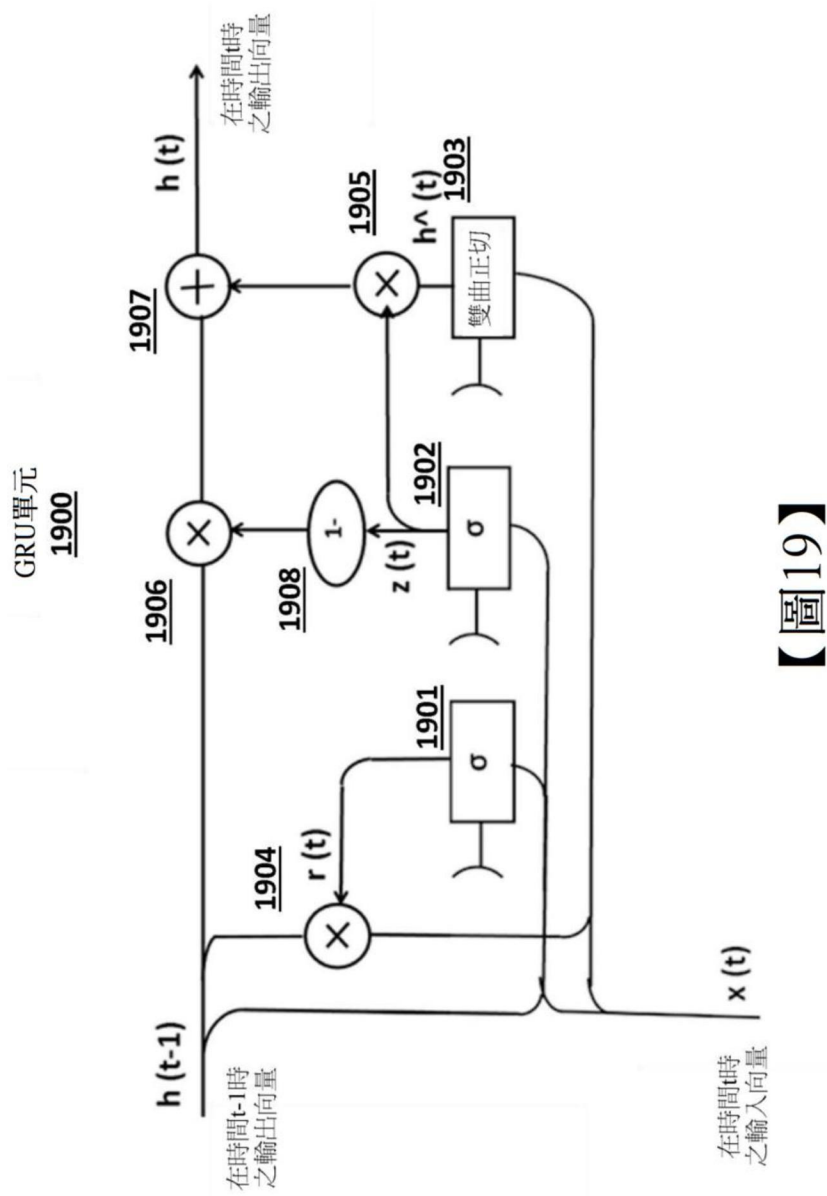


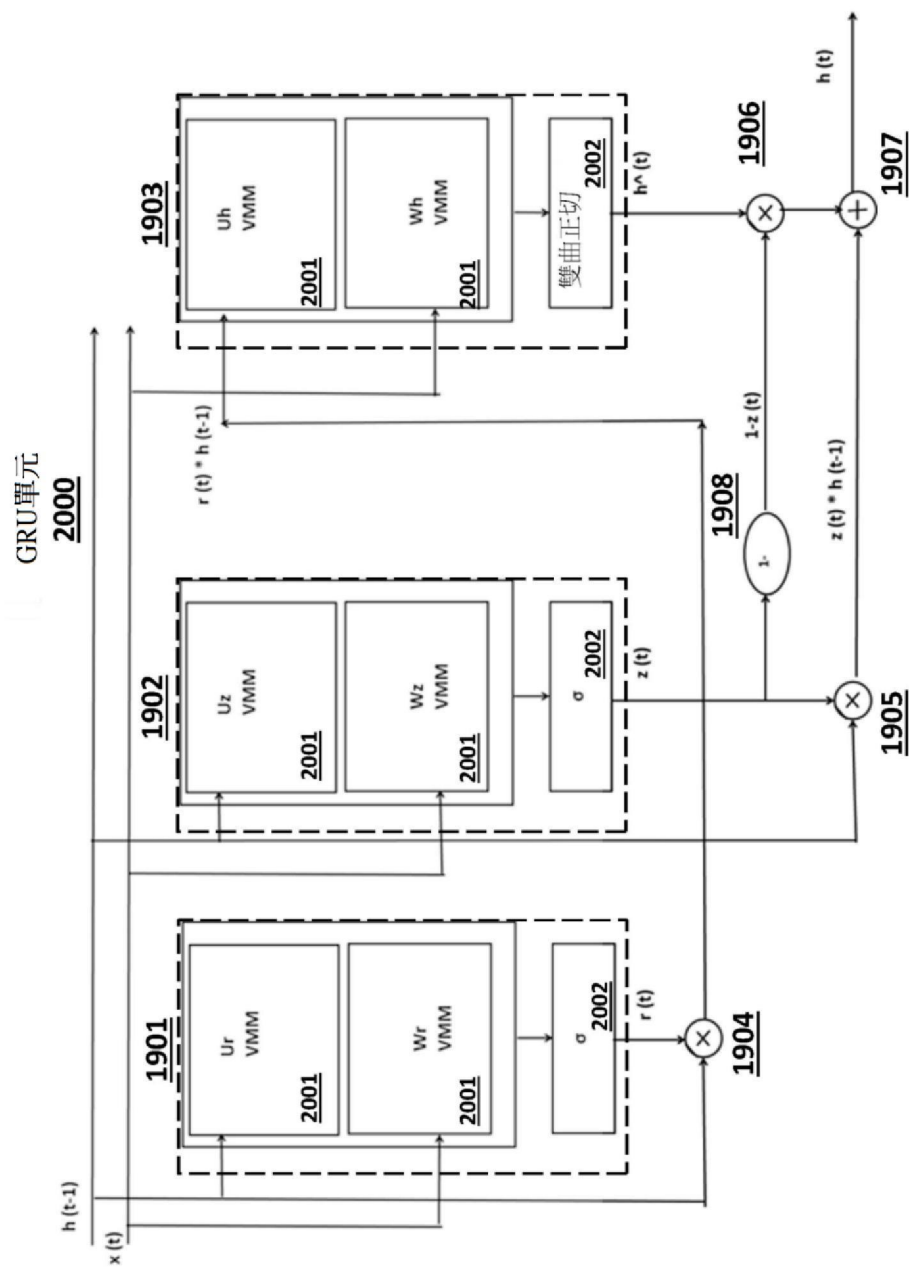
【圖17】

GRU
1800

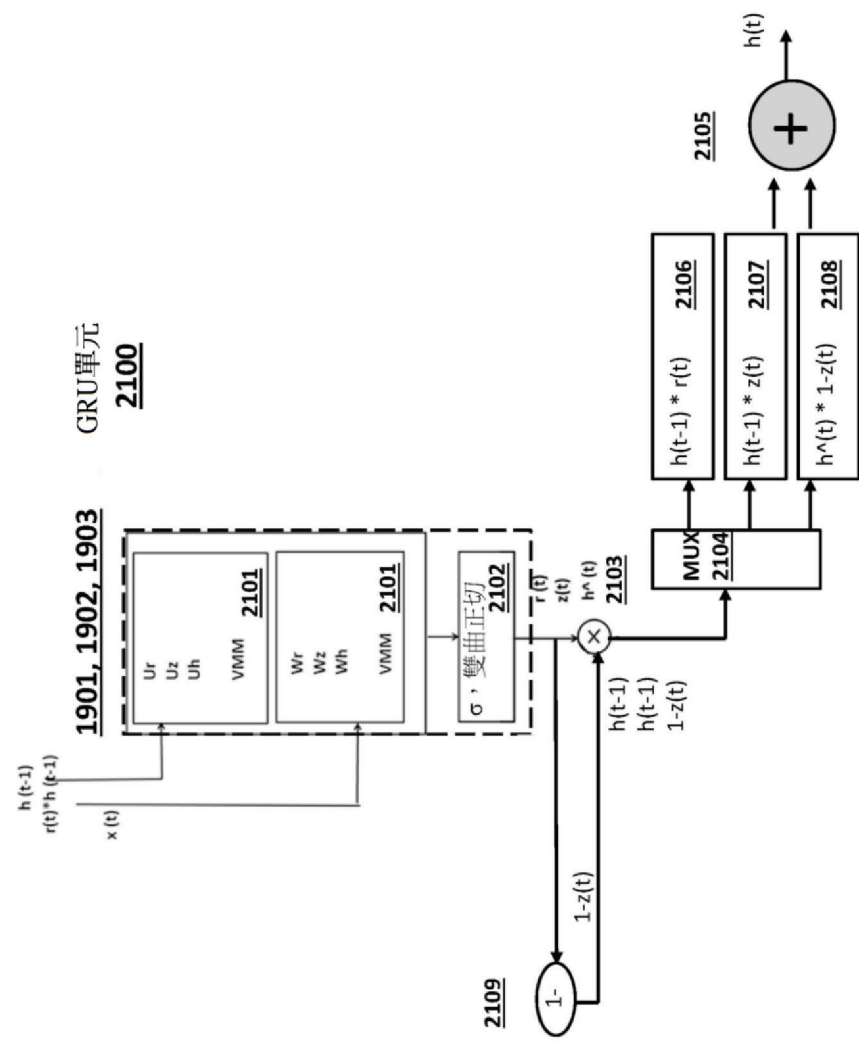


【圖18】

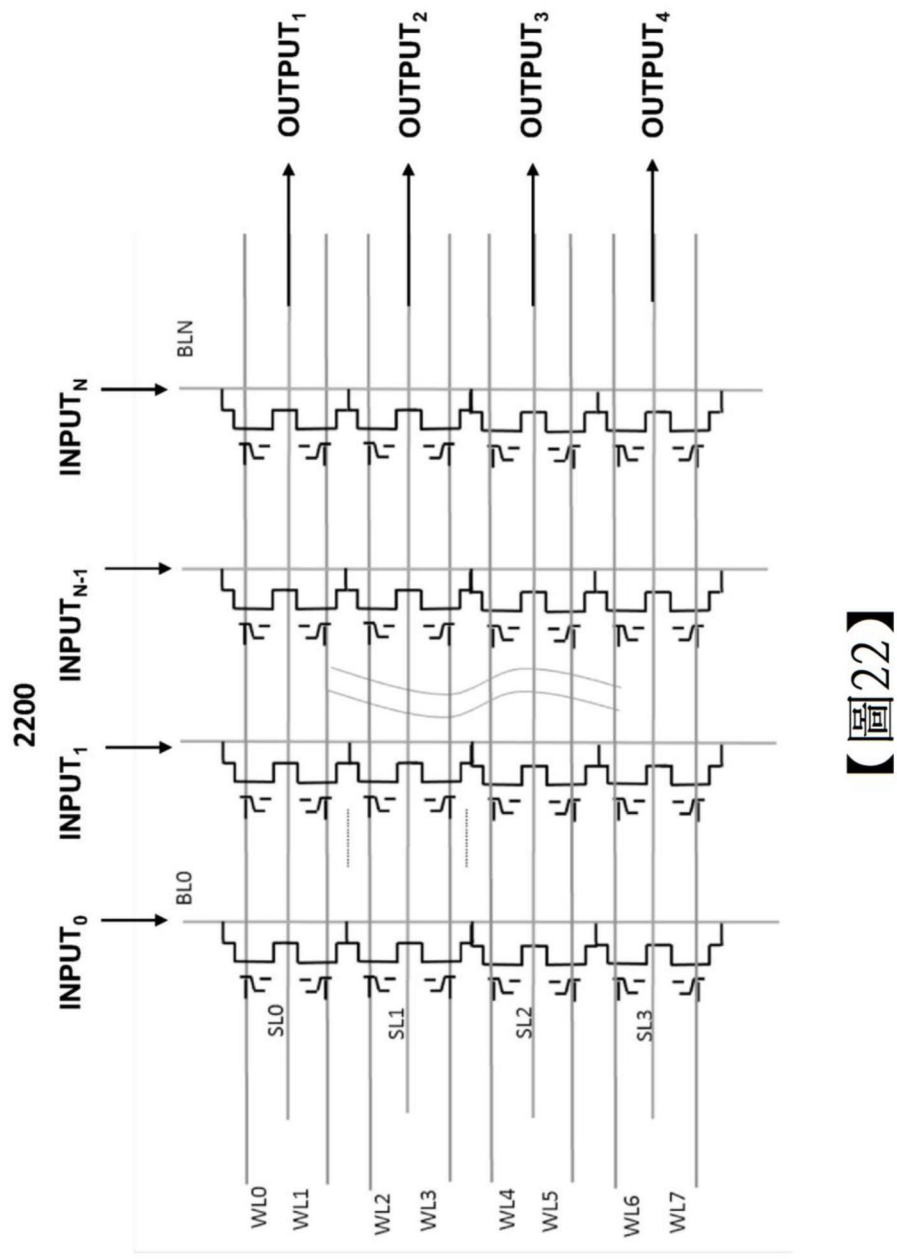




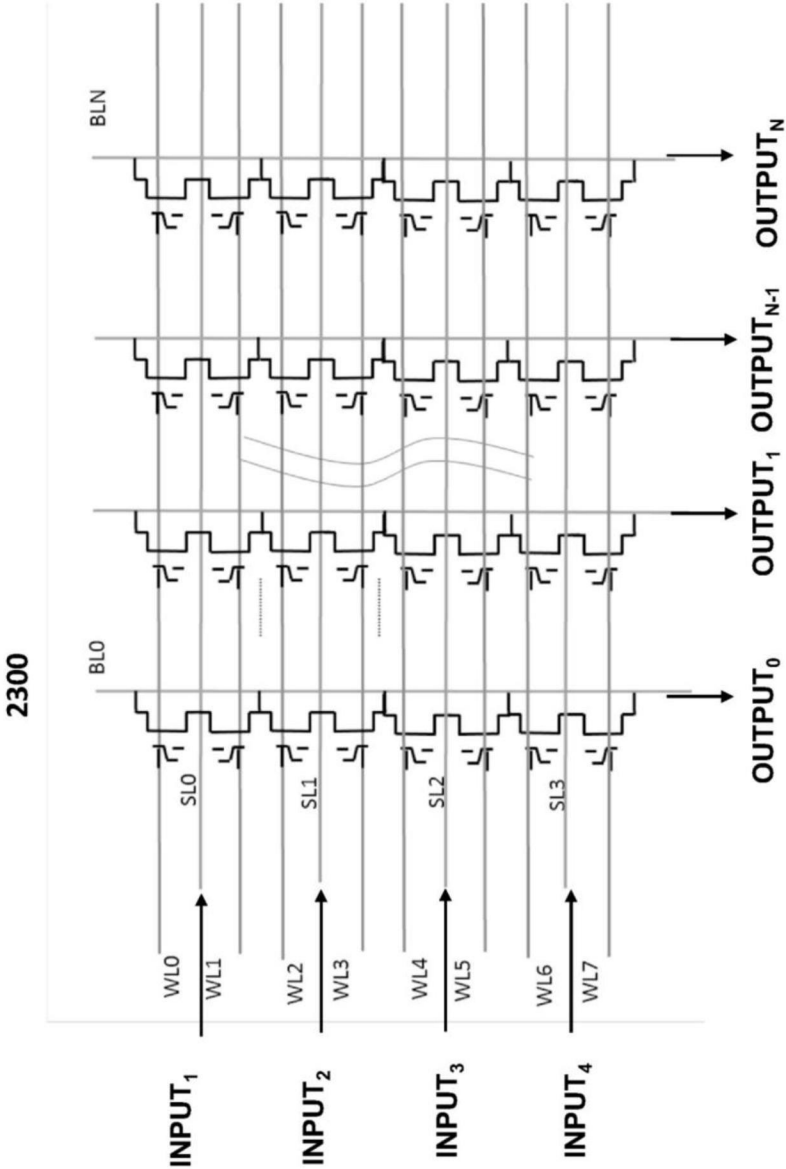
【圖20】



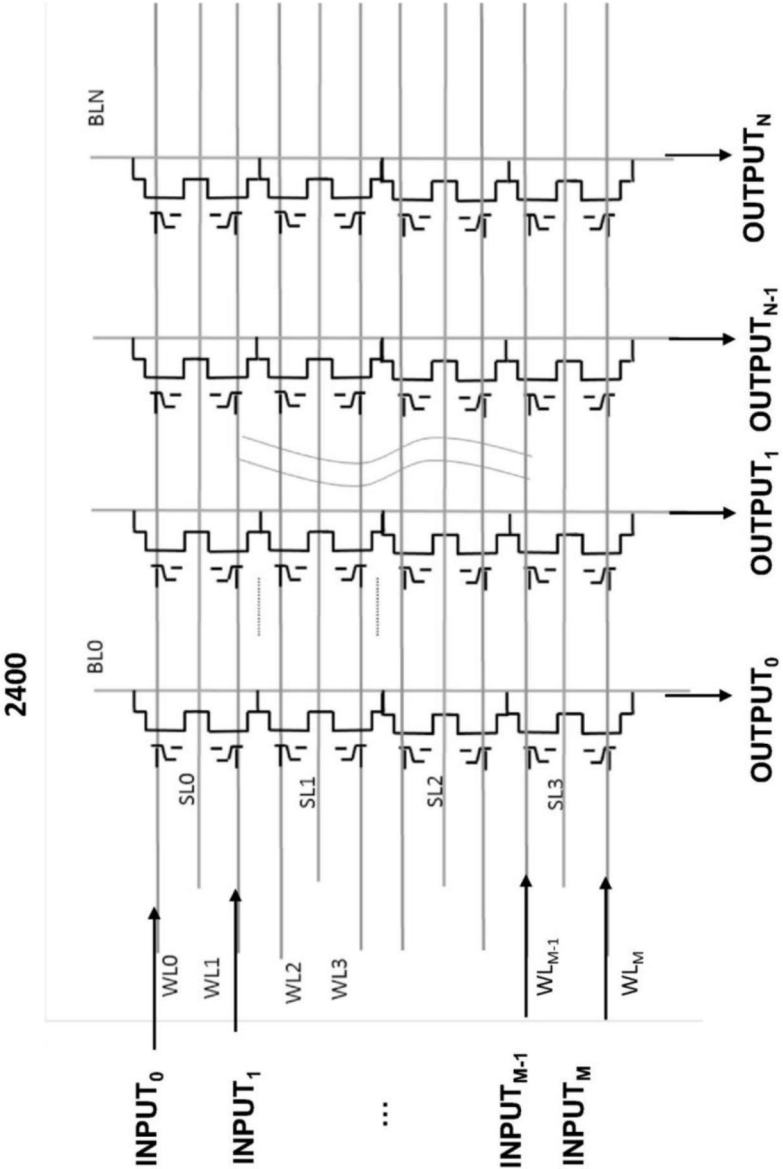
【圖21】



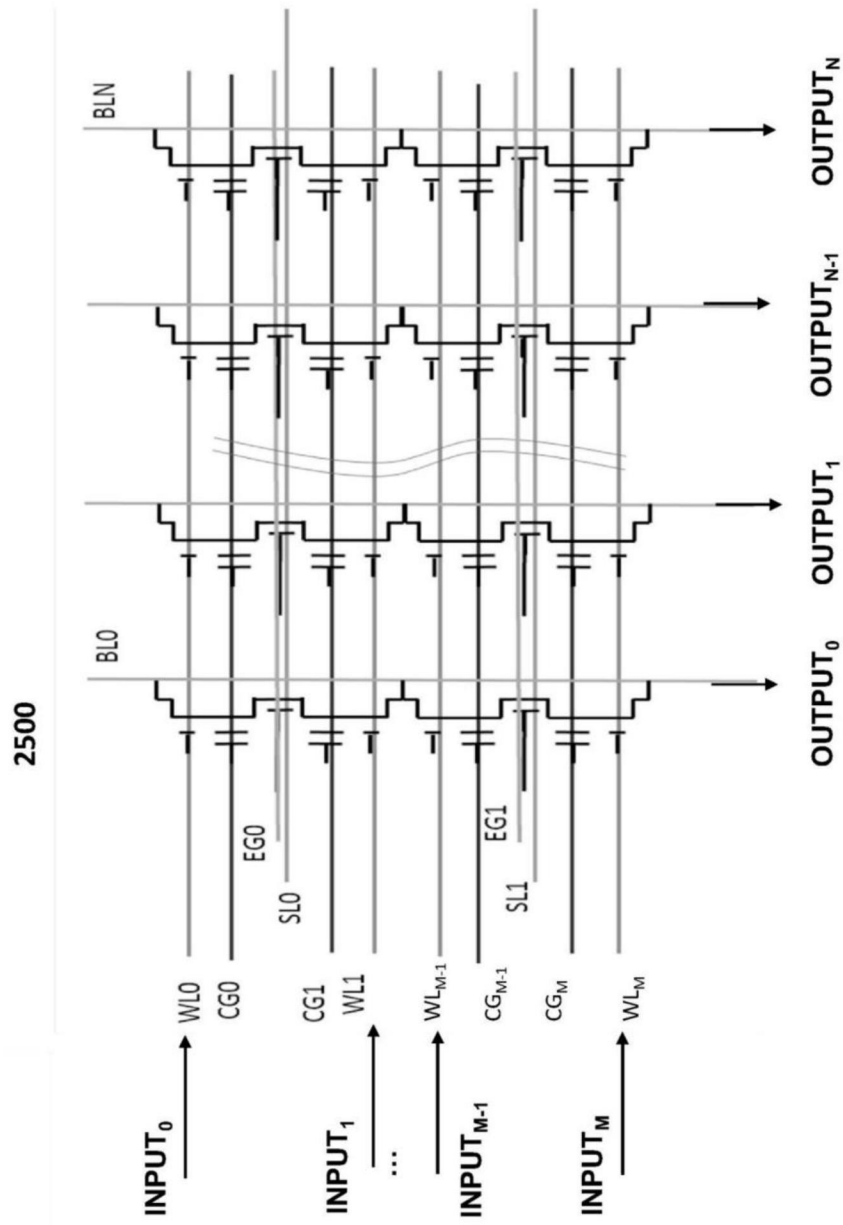
【圖22】



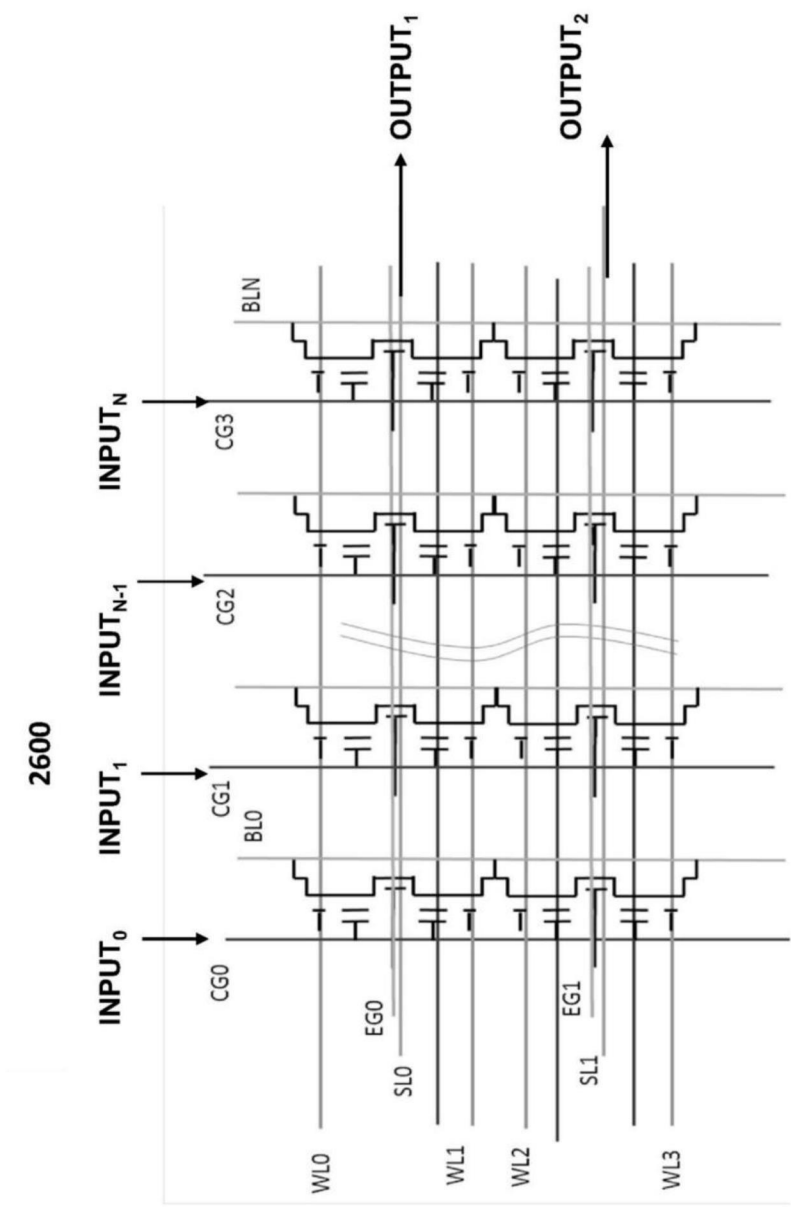
【圖23】



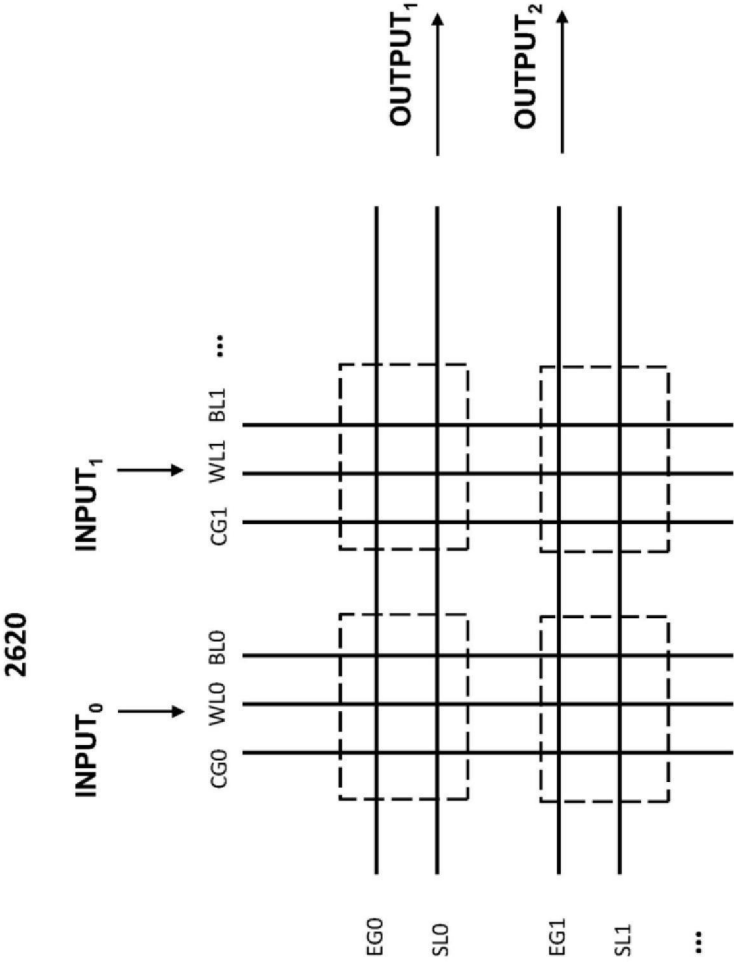
【圖24】



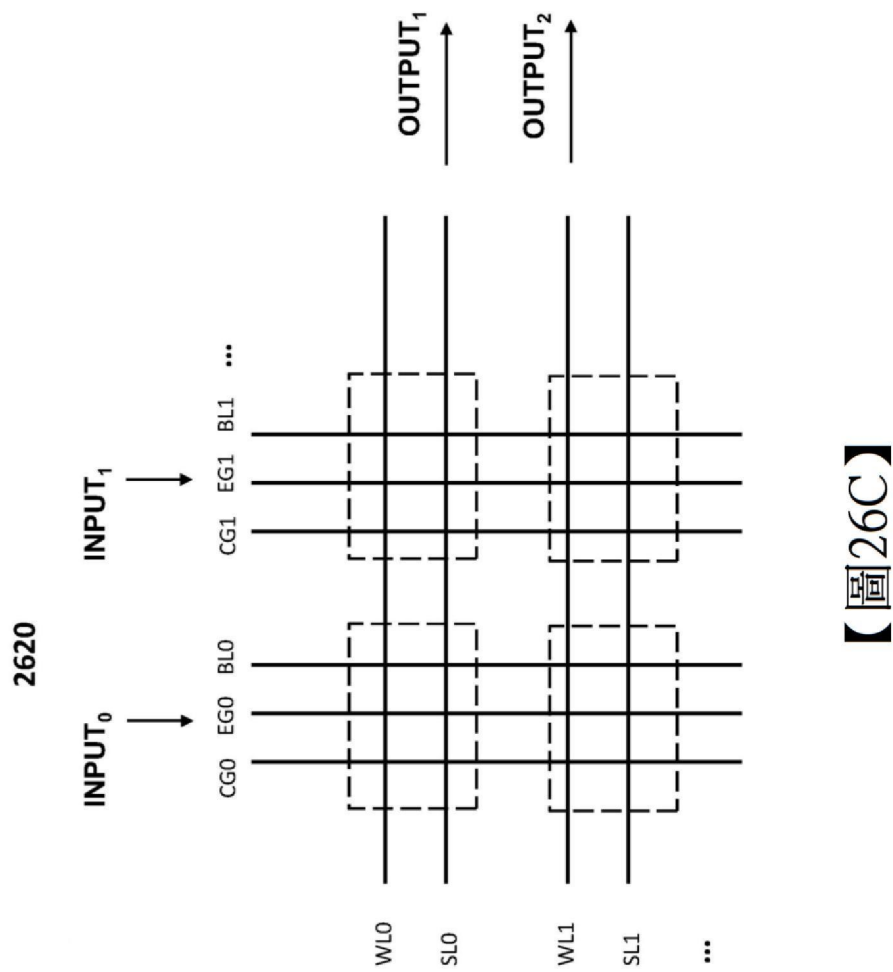
【圖25】



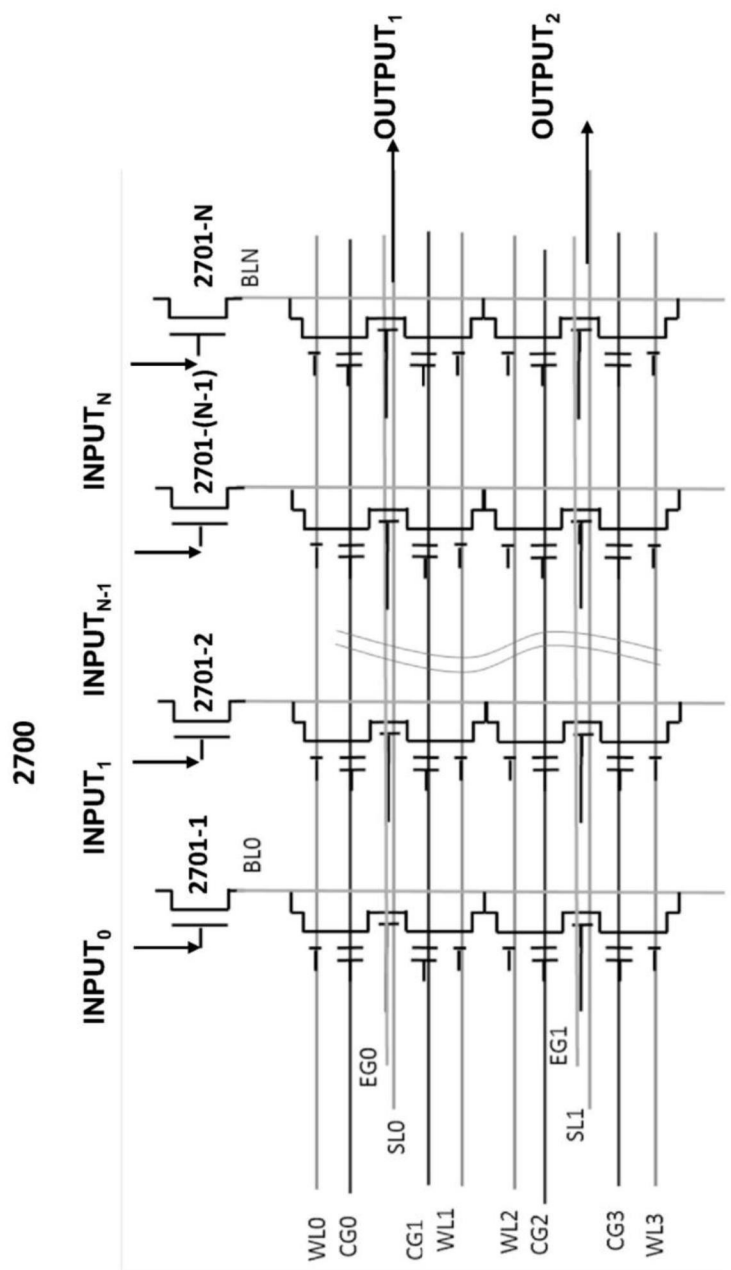
【圖26A】



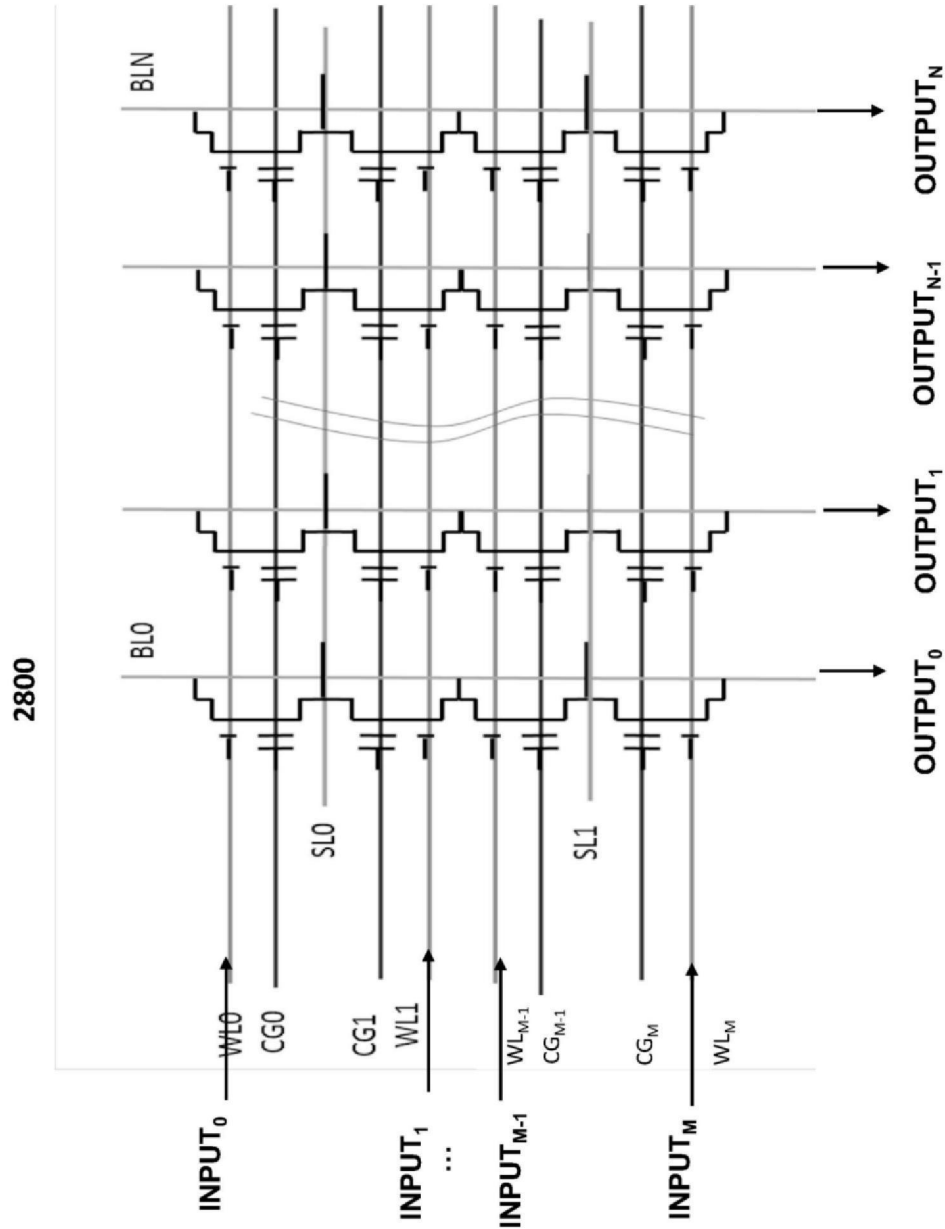
【圖26B】



【圖26C】

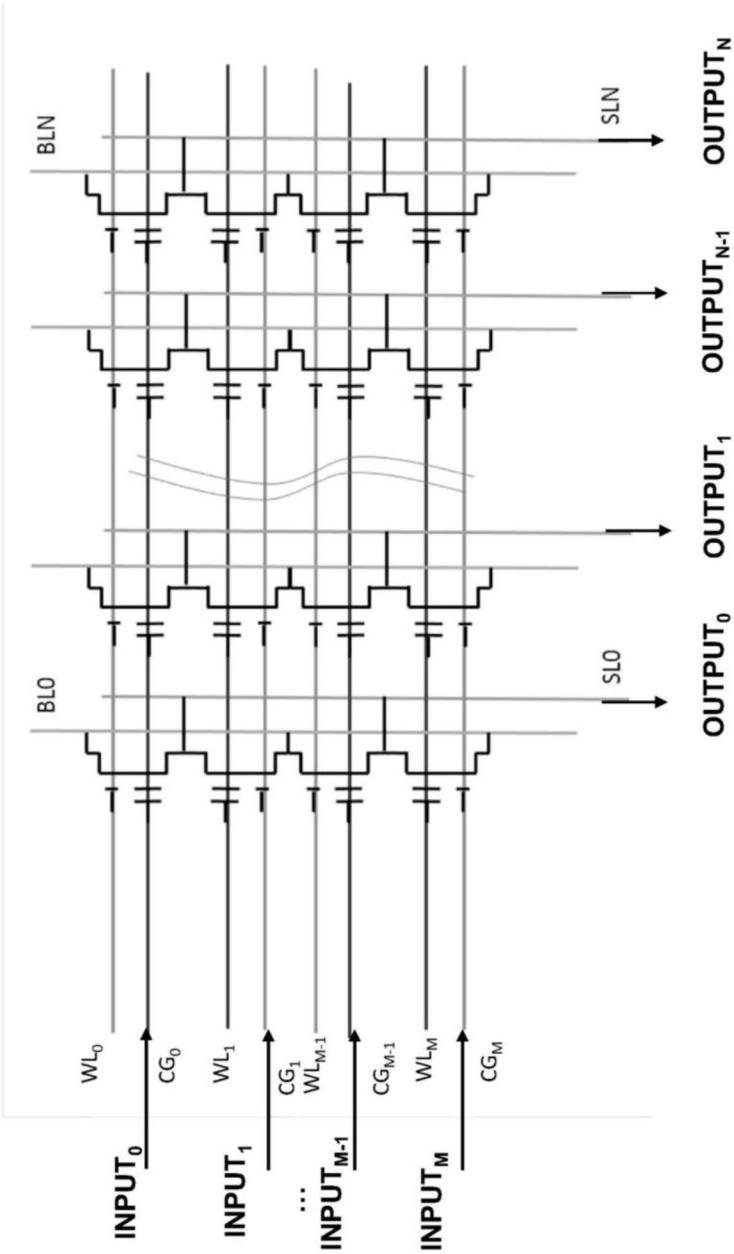


【圖27】

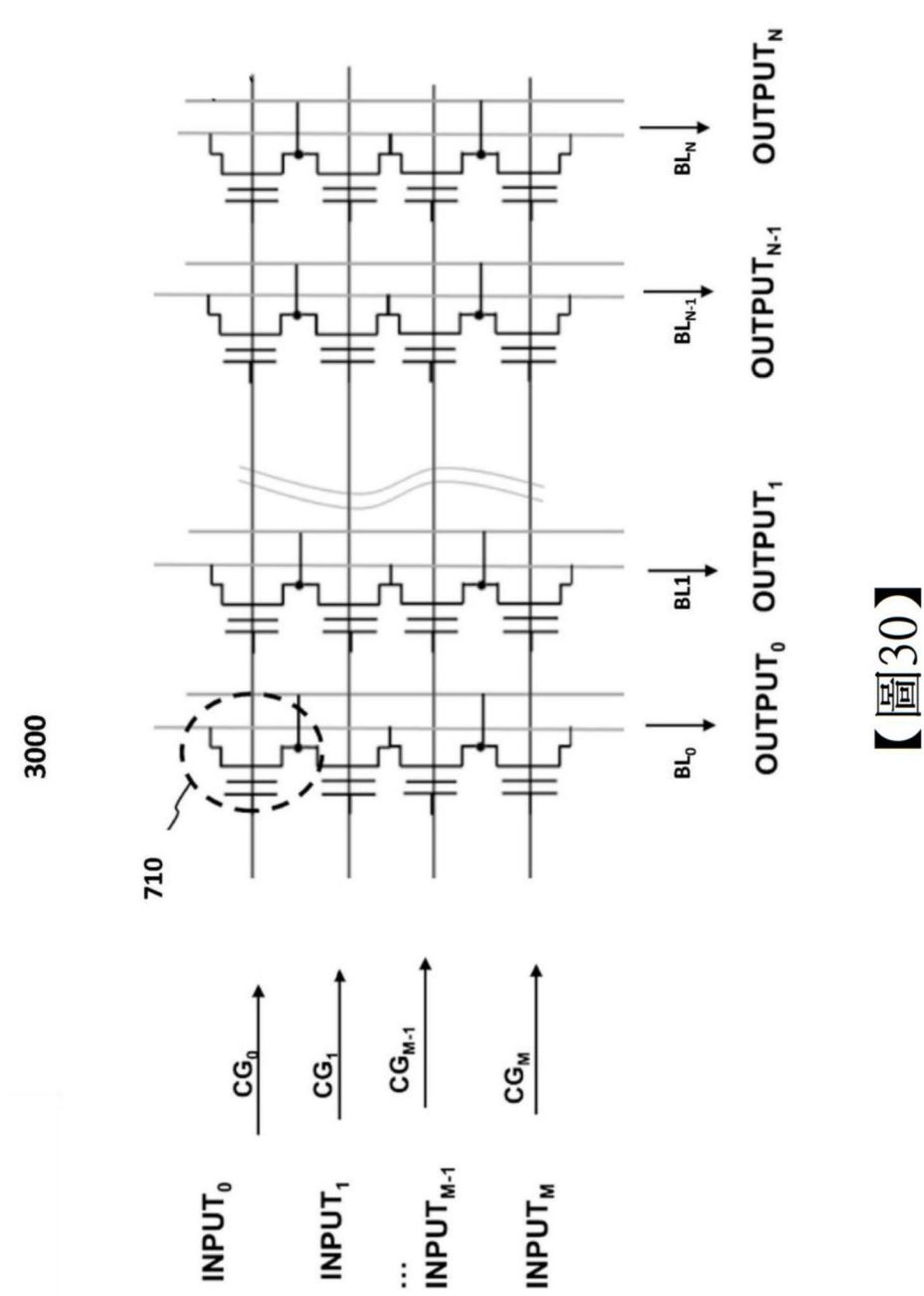


【圖28】

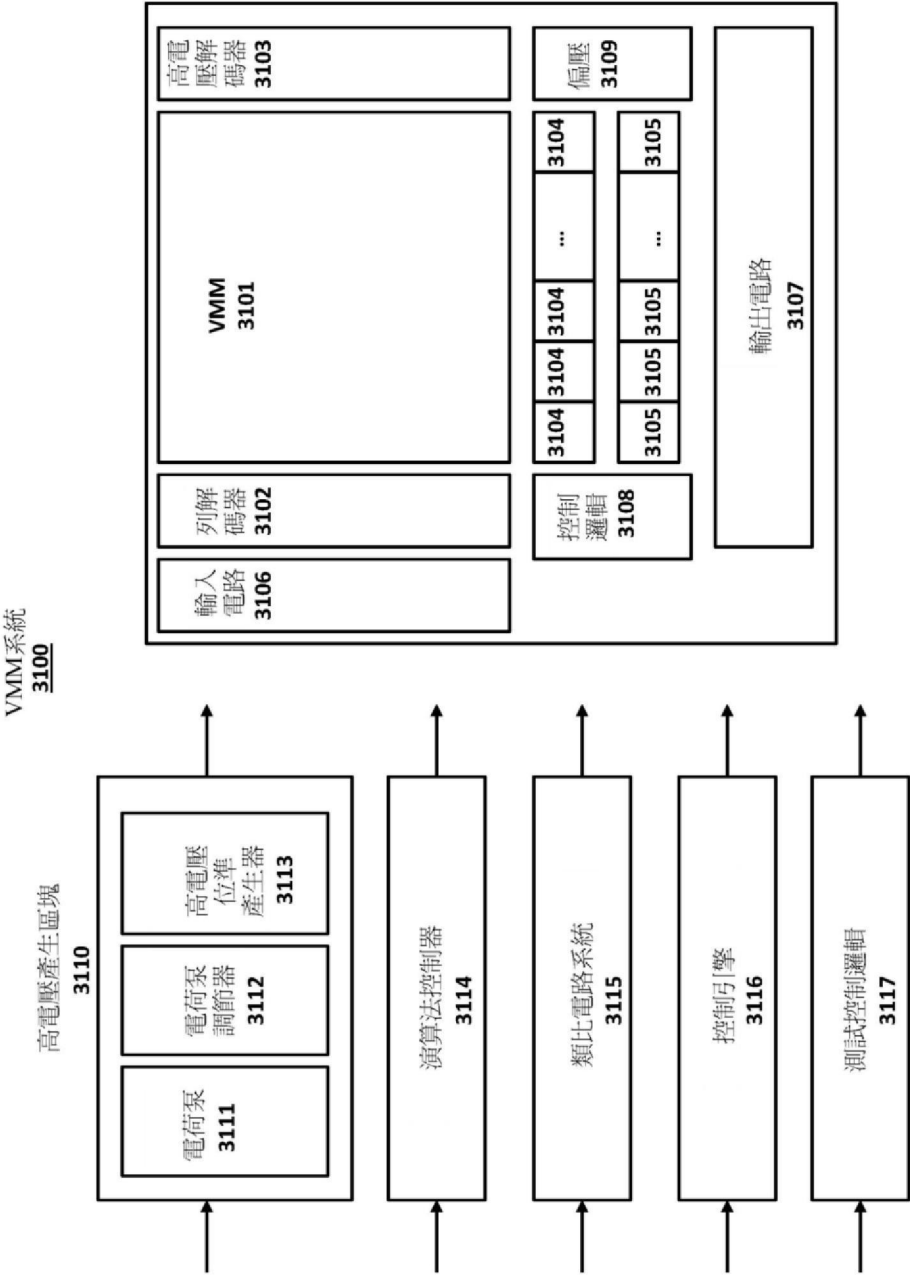
2900



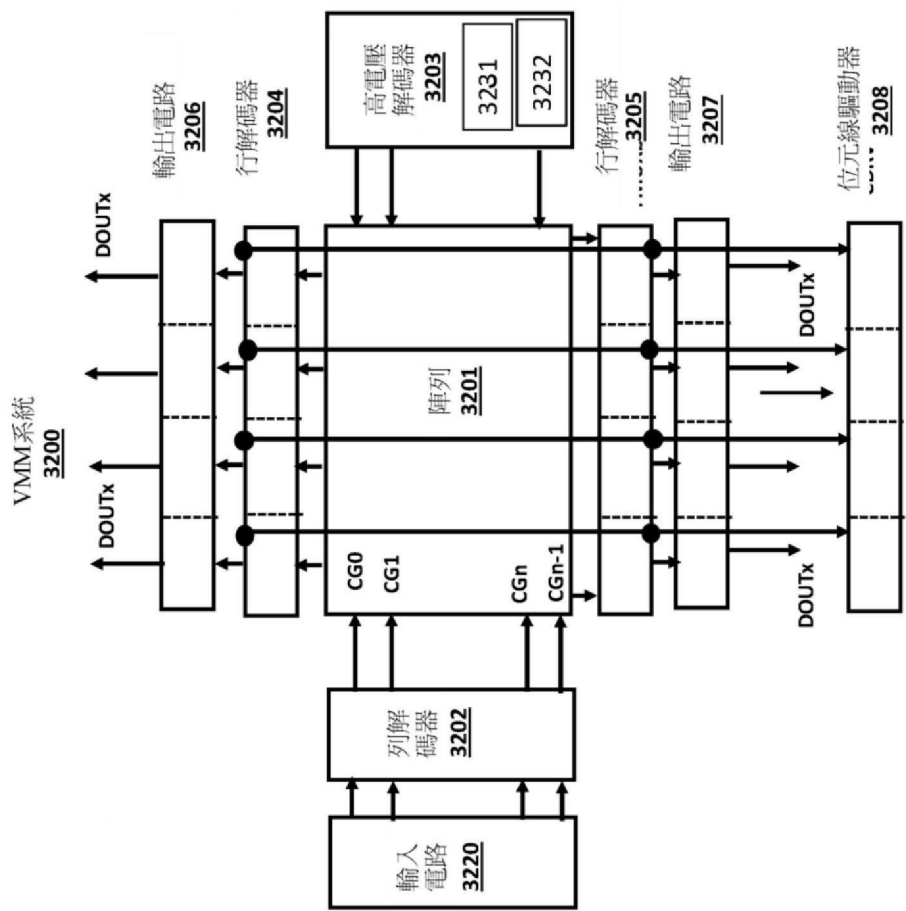
【圖29】



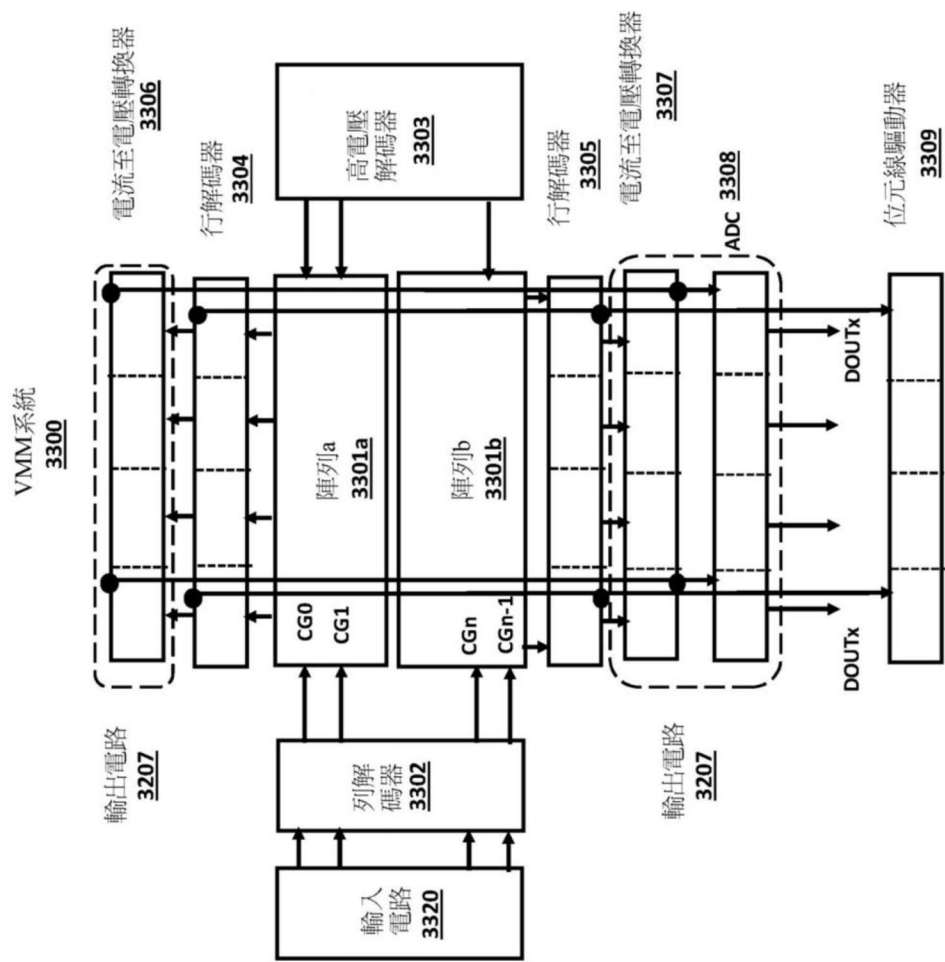
【圖30】



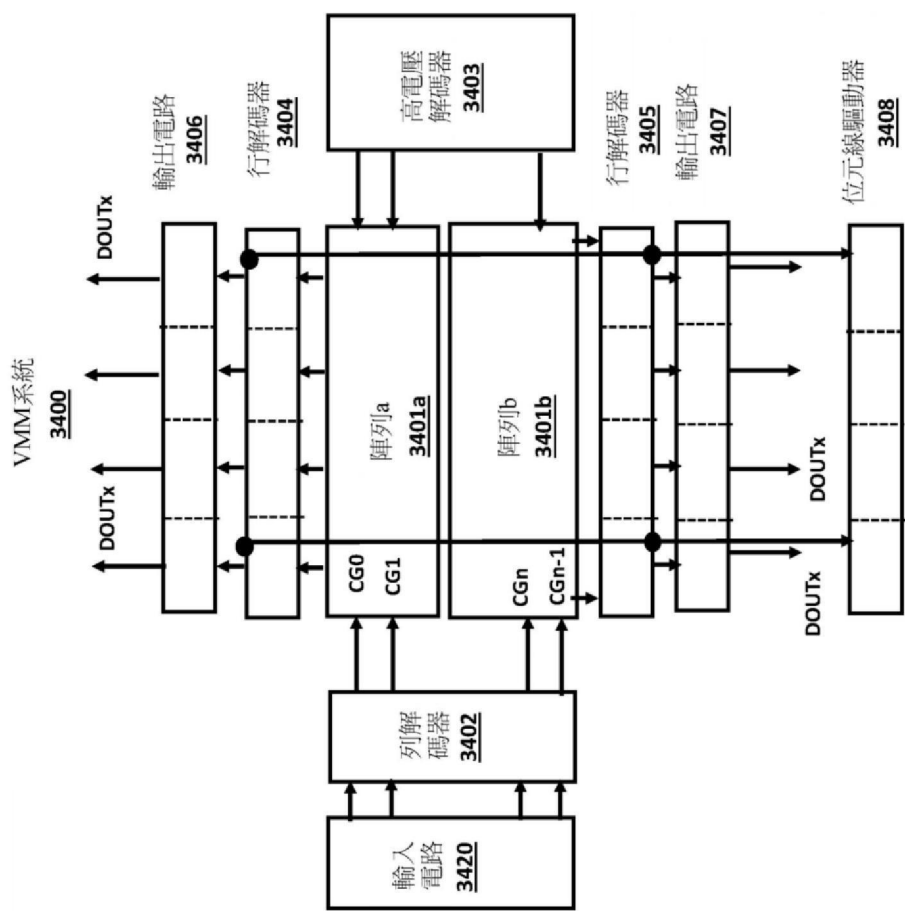
【圖31】



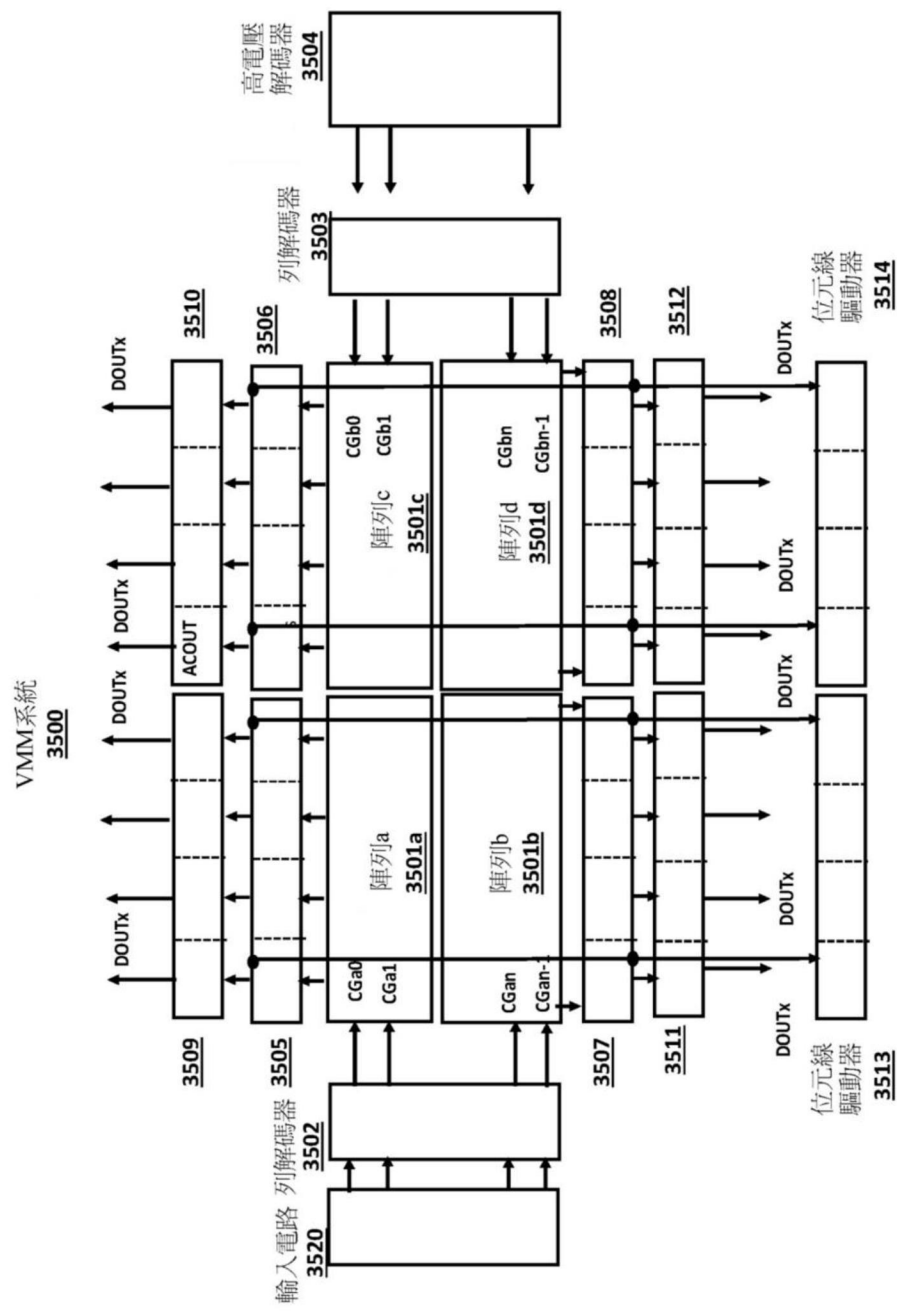
【圖32】



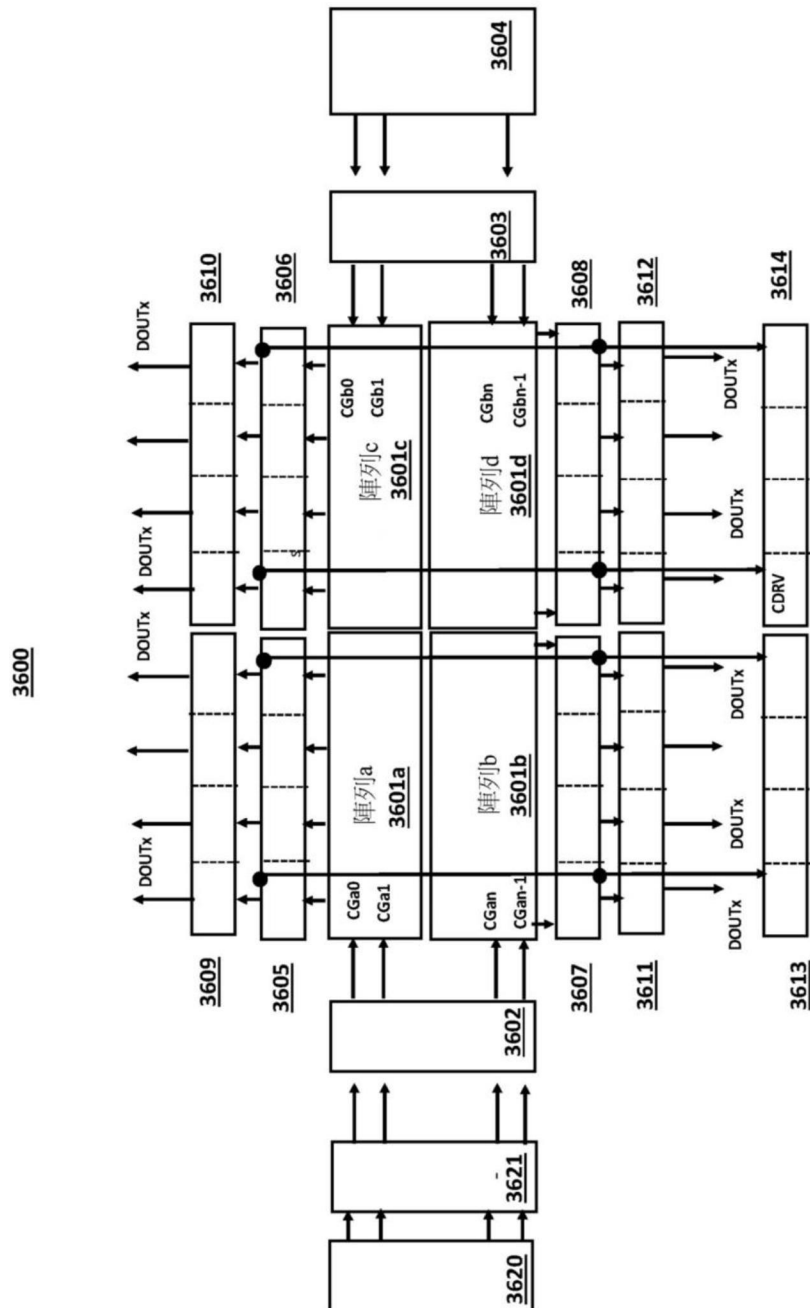
【圖33】



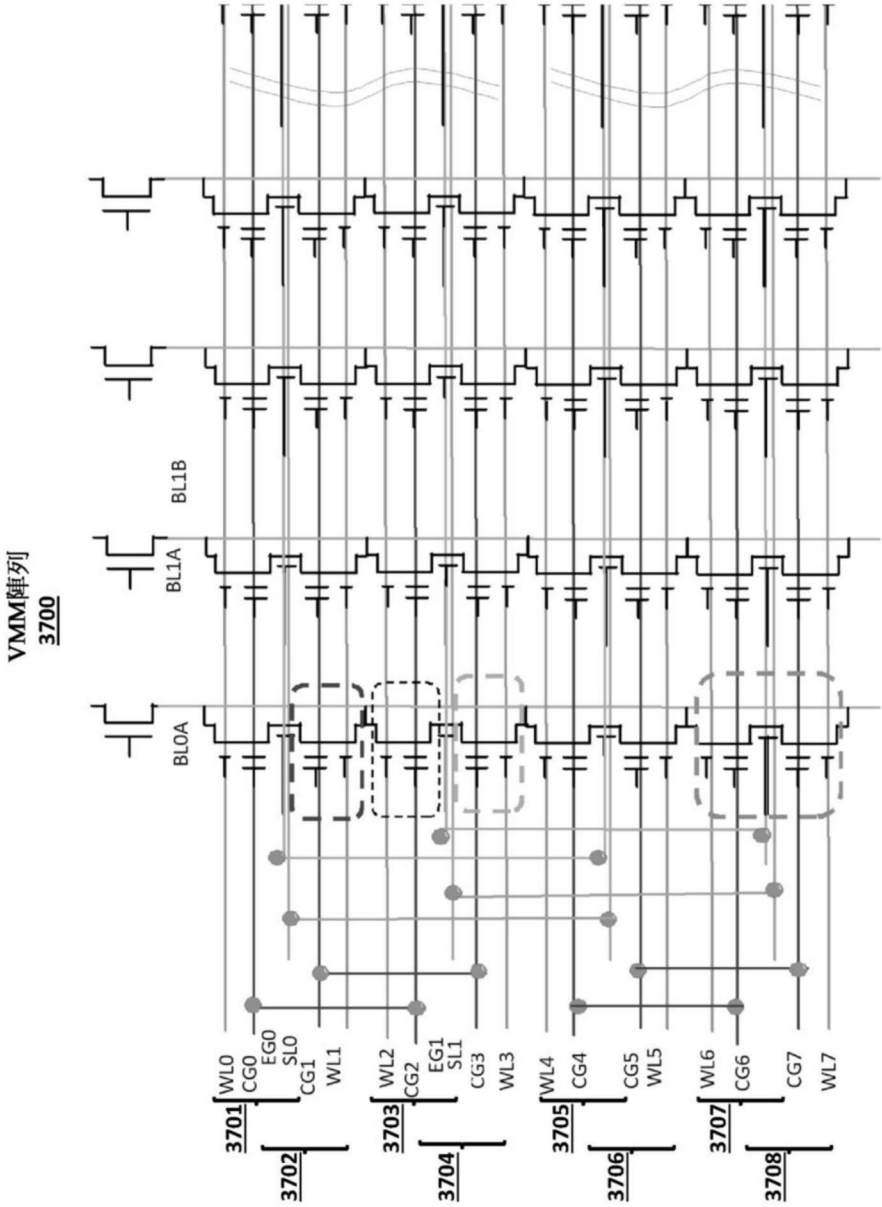
【圖34】



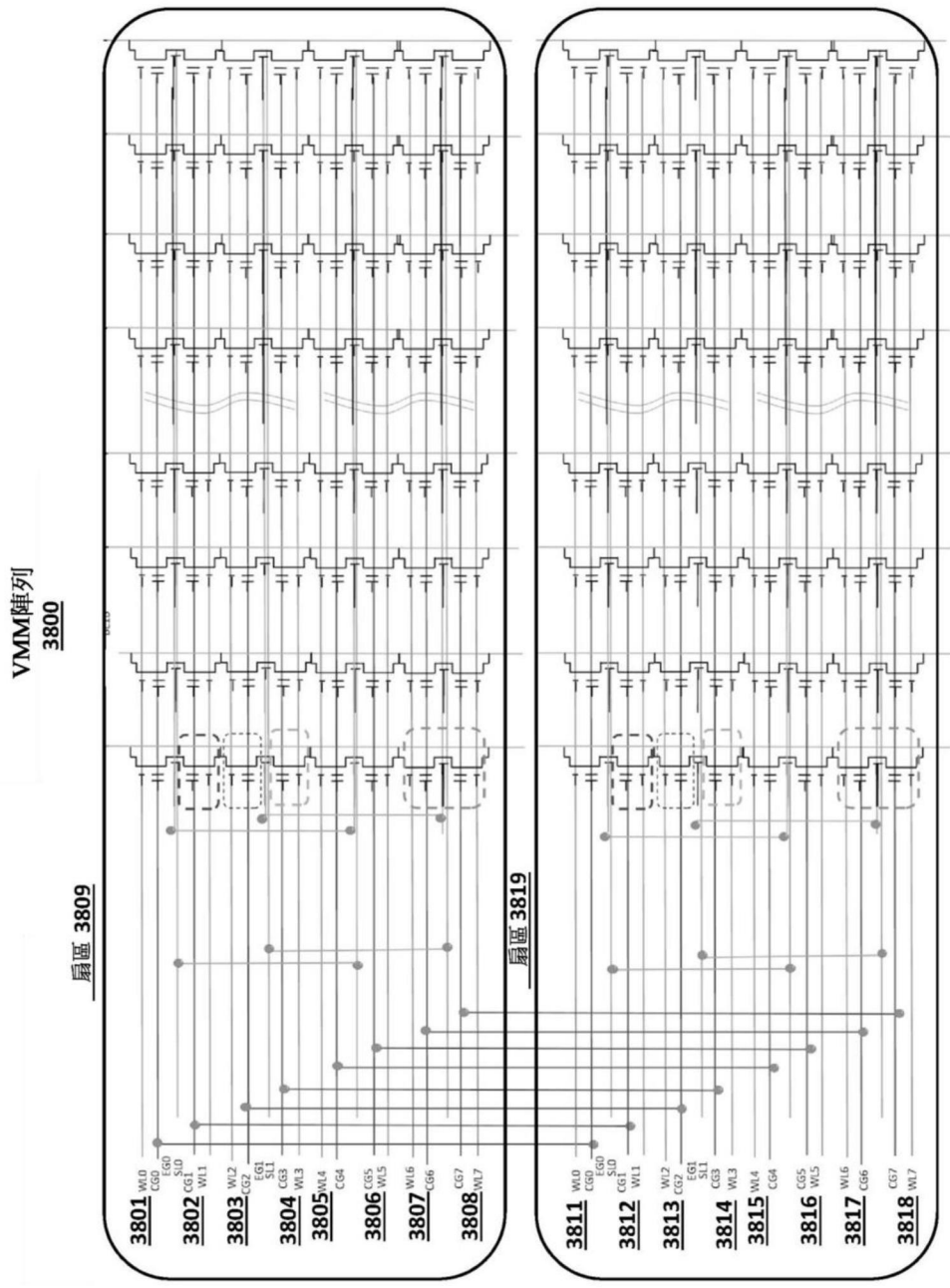
【圖35】



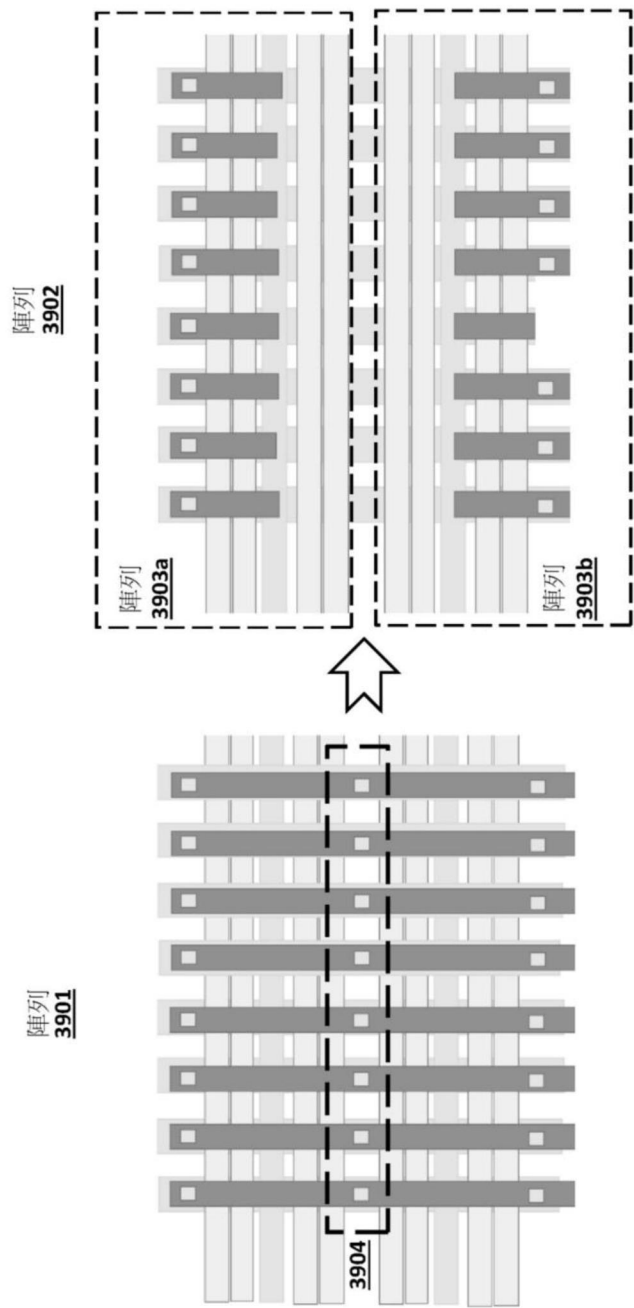
【圖36】



【圖37】



【圖38】



【圖39】