



(21)申請案號：111144915

(22)申請日：中華民國 111 (2022) 年 11 月 24 日

(51)Int. Cl. :

G06F30/27 (2020.01)

G11C8/16 (2006.01)

G11C16/08 (2006.01)

(30)優先權：2022/01/28美國63/304,485

2022/04/14美國17/721,254

2022/04/29世界智慧財產權組織PCT/US22/27046

(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：

TWI694448BUS10741568B2

US2019/0236445A1US2020/0402392A1

審查人員：李惟任

申請專利範圍項數：29 項圖式數：43共 109 頁

(54)名稱

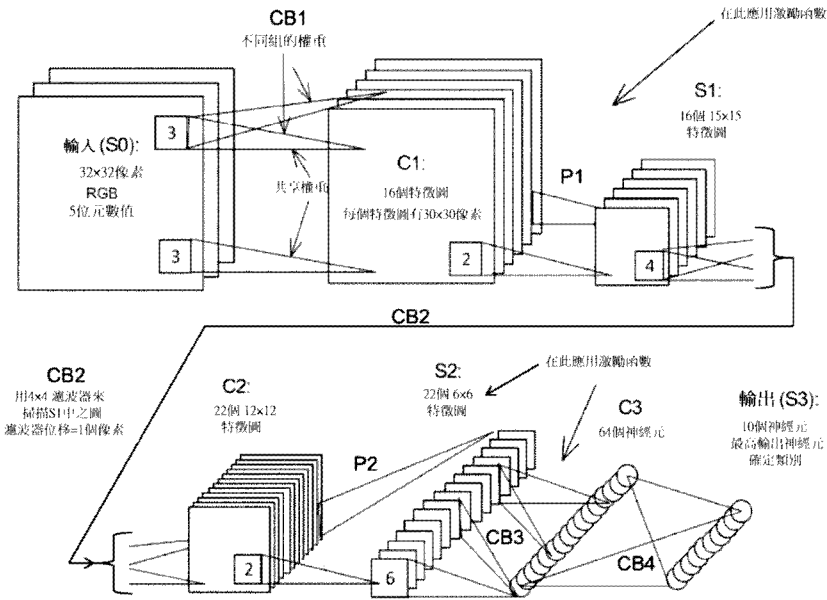
包含類比陣列及數位陣列的人工神經網路

(57)摘要

描述用於提供包含類比陣列及數位陣列之人工神經網路系統的許多實例。在某些實例中，類比陣列及數位陣列耦接至共享位元線。在其它實例中，類比陣列及數位陣列耦接至個別的位元線。

Numerous examples are described for providing an artificial neural network system comprising an analog array and a digital array. In certain examples, an analog array and a digital array are coupled to shared bit lines. In other examples, an analog array and a digital array are coupled to separate bit lines.

指定代表圖：



【圖6】

符號簡單說明：

- C1:層
- C2:層
- C3:層
- CB1:突觸
- CB2:突觸
- CB3:突觸
- CB4:突觸
- P1:激勵函數
- P2:激勵函數
- S0:輸入層
- S1:層
- S2:層
- S3:輸出層



I834397

【發明摘要】

【中文發明名稱】 包含類比陣列及數位陣列的人工神經網路

【英文發明名稱】 ARTIFICIAL NEURAL NETWORK

COMPRISING AN ANALOG ARRAY AND A DIGITAL ARRAY

【中文】

描述用於提供包含類比陣列及數位陣列之人工神經網路系統的許多實例。在某些實例中，類比陣列及數位陣列耦接至共享位元線。在其它實例中，類比陣列及數位陣列耦接至個別的位元線。

【英文】

Numerous examples are described for providing an artificial neural network system comprising an analog array and a digital array. In certain examples, an analog array and a digital array are coupled to shared bit lines. In other examples, an analog array and a digital array are coupled to separate bit lines.

【指定代表圖】 圖6

【代表圖之符號簡單說明】

C1:層

C2:層

C3:層

CB1:突觸

CB2:突觸

CB3:突觸

CB4:突觸

P1:激勵函數

P2:激勵函數

S0:輸入層

S1:層

S2:層

S3:輸出層

【發明說明書】

【中文發明名稱】 包含類比陣列及數位陣列的人工神經網路

【英文發明名稱】 ARTIFICIAL NEURAL NETWORK

COMPRISING AN ANALOG ARRAY AND A DIGITAL ARRAY

【技術領域】

【0001】 [優先權請求]本申請案主張2022年1月28日所提出之名稱為「包含數位資訊陣列之人工神經網路」的美國臨時專利申請案第63/304,485號及2022年4月14日所提出之名稱為「包含類比陣列及數位陣列之人工神經網路」的美國專利申請案第17/721,254號之優先權。

【0002】 揭露包含類比陣列及數位陣列之人工神經網路的許多實例。

【先前技術】

【0003】 人工神經網路模擬生物神經網路(動物的中樞神經系統，特別是大腦)及用於估計或近似可依賴於大量輸入且通常是未知的函數。人工神經網路通常包括可彼此交換信息之互連的「神經元」層。

【0004】 圖1繪示人工神經網路，其中圓圈表示神經元輸入或層。連結(稱為突觸)以箭頭來表示，並且具有可根據經驗調整的數字權重。這使得神經網路適應於輸入且能夠學習。通常，神經網路包括一層多個輸入。通常有一個或多個神經元中間層及提供神經網路輸出的一個神經元輸出層。每個層級的神經元個別地或共同地根據從突觸接收的資料做出決定。

【0005】 開發用於高性能資訊處理之人工神經網路的其中一個主要挑戰是缺乏足夠的硬體技術。確實，實際的神經網路依賴於非常大量的突觸，以使神經元之間的高連結性(亦即，非常高的計算並行性)成為可能。原則上，這樣的複雜性可以用數位超級電腦或專用圖形處理單元叢集來實現。然而，除了高成本之外，相較於生物網路，這些方法還因平庸的能量效率而更糟，其中生物網路消耗非常少的能量，主要是因為它們執行低精度類比計算。CMOS類比電路已經用於人工神經網路，但是有鑑於大量的神經元及突觸，大多數CMOS實施的突觸過於龐大。

【0006】 申請人以前在美國專利申請案公開第2017/0337466A1號中揭露一種人工(類比)神經網路，其利用一個或多個非揮發性記憶體陣列作為突觸，在此以提及方式將上述美國專利申請案併入本文。非揮發性記憶體陣列作為類比神經記憶體來操作且包括以列與行排列的非揮發性記憶體單元。神經網路包括複數個第一突觸，其構造成接收複數個第一輸入並由此產生複數個第一輸出；以及複數個第一神經元，其構造成接收複數個第一輸出。複數個第一突觸包括複數個記憶體單元，其中每個記憶體單元包括間隔開的源極及汲極區域，其形成在半導體基板中，並且具有通道區域在其間延伸；浮動閘極，其設置在通道區域的第一部分上方且與通道區域的第一部分絕緣；以及非浮動閘極，其設置在通道區域的第二部分上方且與通道區域的第二部分絕緣。複數個記憶體單元中之每一者儲存與浮動閘極上之一些電子相對應的權重值。複數個記憶體單元將複數個第一輸入乘以儲存的權重值，以產生複數個第一輸出。

非揮發性記憶體單元

【0007】 非揮發性記憶體係眾所周知的。例如，美國專利第 5,029,130 號(「'130 專利」)揭露一種分離式閘極非揮發性記憶體單元(快閃記憶體單元)陣列，並且在此以提及方式將其併入本文。這樣的記憶體單元 210 顯示在圖 2 中。每個記憶體單元 210 包括在半導體基板 12 中形成之源極區域 14 及汲極區域 16，並且在其間具有通道區域 18。浮動閘極 20 形成在通道區域 18 的第一部分上方且與其絕緣(並控制其導電性)，並且形成在源極區域 14 的一部分上方。字元線端子 22(通常耦接至字元線)具有第一部分及第二部分，其中第一部分設置在通道區域 18 的第二部分上方且與其絕緣(並控制其導電性)，而第二部分向上延伸且在浮動閘極 20 上方。浮動閘極 20 及字元線端子 22 藉由閘極氧化物與基板 12 絕緣。位元線 24 耦接至汲極區域 16。

【0008】 藉由在字元線端子 22 上施加高正電壓來抹除記憶體單元 210(其中從浮動閘極移除電子)，這導致浮動閘極 20 上的電子藉由富爾-諾罕穿隧(Fowler-Nordheim tunneling)從浮動閘極 20 隧穿中間絕緣體至字元線端子 22。

【0009】 藉由在字元線端子 22 上施加正電壓及在源極區域 14 上施加正電壓，利用熱電子透過源極側注入(SSI)來程式化記憶體單元 210(其中在浮動閘極上放置電子)。電子流將從汲極區域 16 流向源極區域 14。當電子到達字元線端子 22 與浮動閘極 20 之間の間隙時，電子將加速並變熱。由於來自浮動閘極 20 的靜電吸引力，一些加熱的電子將通過閘極氧化物注入至浮動閘極 20 上。

【0010】 藉由在汲極區域16及字元線端子22上施加正讀取電壓(這會導通在字元線端子下方之通道區域18的部分)來讀取記憶體單元210。如果浮動閘極20帶正電(亦即，被抹除電子)，則浮動閘極20下方之通道區域18的部分亦導通，並且電流將流過通道區域18，因而被感測為抹除狀態或狀態「1」。如果浮動閘極20帶負電(亦即，用電子來程式化)，則浮動閘極20下方之通道區域的部分大部分或完全截止，並且電流不會流過(或者幾乎不流過)通道區域18，因而被感測為程式化狀態或狀態「0」。

【0011】 表1描繪可以施加至記憶體單元210的端子以執行讀取、抹除及程式化操作的典型電壓及電流範圍：

表1：圖2的快閃記憶體單元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】 其它分離式閘極記憶體單元組態係其它類型的快閃記憶體單元且係已知的。例如，圖3描繪4閘極記憶體單元310，其包括源極區域14、汲極區域16、在通道區域18的第一部分上方之浮動閘極20、在通道區域18的第二部分上方之選擇閘極22(通常耦接至字元線WL)、在浮動閘極20上方之控制閘極28以及在源極區域14上方之抹除閘極30。這種組態係在美國專利第6,747,310號中描述，為了各種目的以提及方式將其併入本文。這裡，除浮動閘極20外，所有其它閘極皆是非浮動閘極，這意味著它們電連接或可連接至電壓源。藉由將加熱的電子從通道區域18

注入至浮動閘極20上來執行程式化。藉由電子從浮動閘極20隧穿至抹除閘極30來執行抹除。

【0013】 表2描繪可以施加至記憶體單元310的端子以執行讀取、抹除及程式化操作的典型電壓及電流範圍：

表2：圖3的快閃記憶體單元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描繪3閘極記憶體單元410，其為另一種類型的快閃記憶體單元。除記憶體單元410不具有個別的控制閘極外，記憶體單元410與圖3的記憶體單元310相同。除沒有施加控制閘極偏壓外，抹除操作(藉由抹除閘極的使用來抹除)及讀取操作相似於圖3的操作。程式化操作亦在沒有控制閘極偏壓的情況下完成，結果，在程式化操作期間必須在源極線上施加較高電壓，以補償控制閘極偏壓的缺少。

【0015】 表3描繪可以施加至記憶體單元410的端子以執行讀取、抹除及程式化操作的典型電壓及電流範圍：

表3：圖4的快閃記憶體單元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描繪堆疊式閘極記憶體單元510，其為另一種類型的快閃記憶體單元。除浮動閘極20在整個通道區域18上方延伸及控制閘極22(在此將耦接至字元線)在浮動閘極20上方延伸且以絕緣層(未顯示)隔開外，記憶體單元510相似於圖2的記憶體單元210。抹除係藉由電子從FG至基板的FN穿隧來完成，程式化係藉由電子從源極區域14流向汲極區域16及在通道區域18與汲極區域16之間的區域處進行通道熱電子(CHE)注入來完成，以及讀取操作與具有較高控制閘極電壓的記憶體單元210相似。

【0017】 表4描繪可以施加至記憶體單元510的端子及基板12以執行讀取、抹除及程式化操作的典型電壓範圍：

表4：圖5的快閃記憶體單元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】 本文描述的方法及手段可以適用於其它非揮發性記憶體技術，例如，FINFET分離式閘極快閃或堆疊式閘極快閃記憶體、NAND快閃記憶體、SONOS(矽-氧化物-氮化物-氧化物-矽，氮化物中的電荷陷阱)、MONOS(金屬-氧化物-氮化物-氧化物-矽，氮化物中的金屬電荷陷阱)、ReRAM(電阻式RAM)、PCM(相變記憶體)、MRAM(磁性RAM)、FeRAM(鐵電式RAM)、CT(電荷陷阱)記憶體、CN(碳管)記憶體、OTP(雙級或多級一次性可程式)及CeRAM(相關電子RAM)，但不限於此。

【0019】 為了在人工神經網路中利用包括上述其中一種類型的非揮發性記憶體單元的記憶體陣列，實施兩種修改。第一，如下面進一步說明，線路配置成使得每個記憶體單元可以個別地被程式化、抹除及讀取，而不會不利地影響陣列中之其它記憶體單元的記憶狀態。第二，提供記憶體單元的連續(類比)程式化。

【0020】 具體地，陣列中之每個記憶體單元的記憶狀態(亦即，浮動閘極上的電荷)可以獨立地且以對其它記憶體單元的最小干擾從完全抹除狀態連續地變成完全程式化狀態，反之亦然。這意味著單元儲存實際上是類比的，或者至少可以儲存許多離散值(例如，16或64個不同值)中之一個，這允許對記憶體陣列中之所有記憶體單元進行非常精確且個別的調整，並且這使記憶體陣列非常適合儲存神經網路的突觸權重及對其進行微調。

使用非揮發性記憶體單元陣列的神經網路

【0021】 圖6概念性地繪示利用目前實例的非揮發性記憶體陣列之神經網路的一個非限制性實例。此實例將非揮發性記憶體陣列神經網路用於臉部辨識應用，但是可以使用以非揮發性記憶體陣列為基礎的神經網路來實施任何其它適當的應用。

【0022】 S0係輸入層，對於這個實例，其為具有5位元精度的32×32像素RGB影像(亦即，三個32×32像素陣列，一個陣列用於各自的顏色R、G及B，每個像素為5位元精度)。從輸入層S0至層C1的突觸CB1在某些情況下應用不同組的權重，而在其它情況下應用共享權重，並且用3×3像素重疊濾波器(核心)掃描輸入影像，將濾波器移位1個像素(或者根

據模型所規定，多於1個像素)。具體地，提供用於影像的一個 3×3 部分中之9個像素的數值(亦即，稱為一個濾波器或核心)給突觸CB1，此處這9個輸入值被乘以適當的權重，並且在計算乘法輸出的總和之後，由CB1的第一突觸確定及提供單一輸出值，以便產生層C1的其中一個特徵圖(feature maps)的一個像素。然後，在輸入層S0內將 3×3 濾波器向右移動一個像素(亦即，添加在右側之三個像素的行及丟棄在左側之三個像素的行)，藉以將這個新定位的濾波器中之9個像素值提供給突觸CB1，此處它們被乘以相同的權重，並且由相關的突觸確定第二個單一輸出值。持續這個過程，直到 3×3 濾波器針對所有三種顏色及所有位元(精度值)掃描輸入層S0的整個 32×32 像素影像為止。然後，使用不同組的權重重複這個過程，以產生C1的一個不同特徵圖，直到已經計算層C1的所有特徵圖為止。

【0023】 在層C1處，在本實例中，具有16個特徵圖，每個特徵圖有 30×30 像素。每個像素是從輸入與核心相乘得到之新特徵像素，因此每個特徵圖是二維陣列，因此在這個實例中，層C1構成16層二維陣列(記住這裡引用的層及陣列是邏輯關係，不一定是實體關係-亦即，陣列不一定以實體二維陣列來定向)。層C1中之16個特徵圖的每個特徵圖由應用於濾波器掃描之16組不同的突觸權重中之一組來產生。C1特徵圖可以全部有關於諸如邊界識別之同一個影像特徵的不同態樣。例如，第一圖(使用第一組權重所產生，第一組權重對用於產生此第一圖的所有掃描係共享的)可以識別圓形邊緣，第二圖(使用與第一組權重不同的第二組權重所產生)可以識別矩形邊緣或某些特徵的縱橫比等等。

【0024】 在從層C1到層S1之前應用激勵函數P1(池化(pooling))，其對來自每個特徵圖中之連續的非重疊 2×2 區域的數值進行池化。池化函數P1的目的是算出附近位置的平均值(或者亦可以使用最大值函數)，以減少例如邊緣位置的依賴性及在進入下一個操作之前減小資料大小。在層S1處，具有16個 15×15 特徵圖(亦即，16個不同陣列，每個陣列有 15×15 像素)。從層S1到層C2的突觸CB2用 4×4 濾波器掃描層S1中之圖，並且有一個像素的濾波器移位。在層C2處，具有22個 12×12 特徵圖。在從層C2到層S2之前應用激勵函數P2(池化)，其對來自每個特徵圖中之連續非重疊 2×2 區域的數值進行池化。在層S2處，具有22個 6×6 特徵圖。在從層S2到層C3的突觸CB3處應用激勵函數(池化)，其中層C3中之每個神經元經由CB3的個別突觸連接至層S2中之每個圖。在層C3處，具有64個神經元。從層C3到輸出層S3的突觸CB4將C3完全連接至S3，亦即，層C3中之每個神經元連接至層S3中之每個神經元。S3處的輸出包括10個神經元，其中最高輸出神經元確定類別。此輸出可能例如表示原始影像的內容之識別或分類。

【0025】 使用一個陣列的非揮發性記憶體單元或一個陣列的非揮發性記憶體單元之一部分來實施每層突觸。

【0026】 圖7係可用於那個目的之陣列的方塊圖。向量矩陣乘法(VMM)陣列32包括非揮發性記憶體單元，並用作一層與下一層之間的突觸(例如，圖6中之CB1、CB2、CB3及CB4)。具體地，VMM陣列32包括非揮發性記憶體單元陣列33、抹除閘極及字元線閘極解碼器34、控制閘極解碼器35、位元線解碼器36以及源極線解碼器37，它們對非揮發性記

記憶體單元陣列33的個別輸入進行解碼。對VMM陣列32的輸入可以來自抹除閘極及字元線閘極解碼器34或來自控制閘極解碼器35。此實例中之源極線解碼器37亦對非揮發性記憶體單元陣列33的輸出進行解碼。或者，位元線解碼器36可以對非揮發性記憶體單元陣列33的輸出進行解碼。

【0027】 非揮發性記憶體單元陣列33提供兩個用途。第一，它儲存將由VMM陣列32使用的權重。第二，非揮發性記憶體單元陣列33有效地將輸入乘以非揮發性記憶體單元陣列33中所儲存的權重，並且根據輸出線(源極線或位元線)將它們加起來，以產生輸出，所述輸出將是下一層的輸入或最後一層的輸入。藉由執行乘法及加法函數，非揮發性記憶體單元陣列33不需要個別的乘法及加法邏輯電路，並且因原位記憶體計算而亦具功率效率。

【0028】 非揮發性記憶體單元陣列33的輸出被供應至差分加法器(例如，求和運算放大器或求和電流鏡)38，其計算非揮發性記憶體單元陣列33的輸出之總和，以產生用於卷積的單一數值。差分加法器38配置成執行正權重與負權重的總和。

【0029】 然後，差分加法器38之加總的輸出值被供應至激勵函數區塊39，其對輸出進行整流。激勵函數區塊39可以提供sigmoid、tanh或ReLU函數。激勵函數區塊39之經整流的輸出值變成作為下一層(例如，圖6中之C1)之特徵圖的元素，然後應用於下一個突觸，以產生下一個特徵圖層或最後一層。因此，在此實例中，非揮發性記憶體單元陣列33構成複數個突觸(其從先前的神經元層或從諸如影像資料庫的輸入層接收

它們的輸入)，並且求和運算放大器38及激勵函數區塊39構成複數個神經元。

【0030】 圖7中至VMM陣列32的輸入(WL_x、EG_x、CG_x以及任選的BL_x及SL_x)可以是類比位準、二進制位準或數位位元(在這種情況下，提供DAC，以將數位位元轉換為適當的輸入類比位準)，並且輸出可以是類比位準、二進制位準或數位位元(在這種情況下，提供輸出ADC，以將輸出類比位準轉換為數位位元)。

【0031】 圖8係描繪許多層的VMM陣列32之使用的方塊圖，這裡標記為VMM陣列32a、32b、32c、32d及32e。如圖8所示，藉由數位至類比轉換器31將輸入(表示為Input_x)從數位轉換成類比，並提供至輸入VMM陣列32a。經轉換的類比輸入可以是電壓或電流。用於第一層的輸入D/A轉換可以藉由使用函數或LUT(查找表)來完成，其中LUT(查找表)將輸入Input_x映射至用於輸入VMM陣列32a的矩陣乘數之適當類比位準。輸入轉換亦可以藉由類比至類比(A/A)轉換器來完成，以將外部類比輸入轉換成輸入VMM陣列32a的映射類比輸入。

【0032】 由輸入VMM陣列32a產生的輸出作為輸入提供給下一個VMM陣列(隱藏層級1)32b，其轉而產生作為輸入提供給下一個VMM陣列(隱藏層級2)32c的輸出等等。各種層的VMM陣列32充當卷積神經網路(CNN)之不同層的突觸及神經元。每個VMM陣列32a、32b、32c、32d及32e可以是獨立的實體非揮發性記憶體陣列，或者多個VMM陣列可以利用同一個實體非揮發性記憶體陣列的不同部分，或者多個VMM陣列可以利用同一個實體非揮發性記憶體陣列的重疊部分。圖8所示的實例包含

五層(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)。所屬技術領域之具通常技藝人士將理解，這僅僅是示例性的，並且系統反而可以包括多於兩個隱藏層及多於兩個完全連接層。

向量矩陣乘法(VMM)陣列

【0033】 圖9描繪神經元VMM陣列900，其特別適用於圖3中所示之記憶體單元310，並且用作輸入層與下一層之間的突觸及神經元的部件。VMM陣列900包括非揮發性記憶體單元的記憶體陣列901及非揮發性參考記憶體單元的參考陣列902(在所述陣列的上方)。或者，可以在下方放置另一個參考陣列。

【0034】 在VMM陣列900中，諸如控制閘極線903的控制閘極線在垂直方向上延伸(因此，在列方向上的參考陣列902與控制閘極線903正交)，並且諸如抹除閘極線904的抹除閘極線在水平方向延伸。這裡，在控制閘極線(CG0、CG1、CG2、CG3)上提供對VMM陣列900的輸入，而VMM陣列900的輸出出現在源極線(SL0、SL1)上。在一個具體例中，僅使用偶數列，而在另一個具體例中，僅使用奇數列。在每條源極線(SL0、SL1)上之電流執行來自連接至那條特定源極線之記憶體單元的所有電流之求和函數。

【0035】 如本文針對神經網路所述，VMM陣列900的非揮發性記憶體單元(亦即，VMM陣列900的記憶體單元310)操作上係配置成在次臨界區域中操作。

【0036】 在弱倒轉(次臨界區域)中施加偏壓於本文所述之非揮發性參考記憶體單元及非揮發性記憶體單元：

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

$$\text{其中 } w = e^{(-V_{th})/nV_t}$$

其中 I_{ds} 係汲極至源極電流； V_g 係記憶體單元上的閘極電壓； V_{th} 係記憶體單元的臨界電壓； V_t 係熱電壓= $k*T/q$ ，其中 k 係波茲曼常數， T 係克耳文溫度， q 係電子電荷； n 係斜率因子= $1+(C_{dep}/C_{ox})$ ，其中 C_{dep} =空乏層的電容， C_{ox} 係閘極氧化層的電容； I_o 係閘極電壓等於臨界電壓時的記憶體單元電流， I_o 與 $(W_t/L)*u*C_{ox}*(n-1)*V_t^2$ 成正比，其中 u 是載子遷移率， W_t 及 L 分別是記憶體單元的寬度及長度。

【0037】 對於使用記憶體單元(例如，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：

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

其中 w_p 係參考或周邊記憶體單元的 w 。

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

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

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

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

在此， w_a =記憶體陣列中之每個記憶體單元的 w 。

V_{thp} 係周邊記憶體單元的有效臨界電壓， V_{tha} 係主(資料)記憶體單元的有效臨界電壓。注意，電晶體的臨界電壓係基板本體偏壓的函數，

表示為 V_{sb} 的基板本體偏壓可以被調製以補償關於這種溫度的各種狀況。臨界電壓 V_{th} 可表示為：

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

其中 V_{th0} 係具有零基板偏壓的臨界電壓， ϕ_F 係表面電位， γ 係本體效應(body effect)參數。

【0039】 字元線或控制閘極可用以作為用於輸入電壓之記憶體單元的輸入。

【0040】 或者，本文所述之VMM陣列的快閃記憶體單元可以配置成在線性區域中操作：

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

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

表示線性區域中的權重 W 與 $(V_{gs} - V_{th})$ 成正比。

【0041】 字元線或控制閘極或位元線或源極線可用以作為在線性區域中操作之記憶體單元的輸入。位元線或源極線可用以作為記憶體單元的輸出。

【0042】 對於I至V線性轉換器，在線性區域中操作的記憶體單元(例如，參考記憶體單元或周邊記憶體單元)或電晶體可以用於將輸入/輸出電流線性地轉換成輸入/輸出電壓。

【0043】 或者，本文所描述之VMM陣列的記憶體單元可以構成在飽和區中操作：

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

$W \propto (V_{gs} - V_{th})^2$ ，表示權重 W 與 $(V_{gs} - V_{th})^2$ 成正比。

【0044】 字元線、控制閘極或抹除閘極可用以作為在飽和區域中操作之記憶體單元的輸入。位元線或源極線可用以作為輸出神經元的輸出。

【0045】 或者，本文所描述之VMM陣列的記憶體單元可以用在神經網路之每一層或多層的所有區域或其組合(次臨界、線性或飽和)中。

【0046】 圖7的VMM陣列32之其它實例描述於美國專利第10,748,630號中，在此以提及方式將其併入本文。如上面申請案所述，源極線或位元線可用以作為神經元輸出(電流總和輸出)。

【0047】 圖10描繪神經元VMM陣列1000，其特別適用於圖2所示之記憶體單元210，並且用作輸入層與下一層之間的突觸。VMM陣列1000包括非揮發性記憶體單元的記憶體陣列1003、第一非揮發性參考記憶體單元的參考陣列1001及第二非揮發性參考記憶體單元的參考陣列1002。配置在陣列的行方向上之參考陣列1001及1002用於將流入端子BLR0、BLR1、BLR2及BLR3的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係經由多工器1014(部分被描繪)與流入它們的電流輸入以二極體形式連接。將參考單元調整(例如，程式化)至目標參考位準。目標參考位準由參考微型陣列矩陣(未顯示)來提供。

【0048】 記憶體陣列1003提供兩個用途。第一，它在其個別記憶體單元上儲存將被VMM陣列1000使用的權重。第二，記憶體陣列1003有效地將輸入(亦即，被提供至端子BLR0、BLR1、BLR2及BLR3的電流輸入；參考陣列1001及1002將這些電流輸入轉換成輸入電壓，以供應至

字元線WL0、WL1、WL2及WL3)乘以記憶體陣列1003中所儲存之權重，然後將所有結果(記憶體單元電流)相加，以在個別位元線(BL0-BLN)上產生輸出，所述輸出將是下一層的輸入或最後一層的輸入。藉由執行乘法及加法函數，記憶體陣列1003不需要個別的乘法及加法邏輯電路，並且還具有功率效率。這裡，電壓輸入被提供在字元線WL0、WL1、WL2及WL3上，並且輸出在讀取(推理)操作期間出現在位元線BL0-BLN上。在位元線BL0-BLN的每條位元線上之電流執行來自連接至那條特定位元線之所有非揮發性記憶體單元的電流之求和函數。

【0049】 表5描繪VMM陣列1000的操作電壓及電流。表中之行表示在被選單元的字元線、未被選單元的字元線、被選單元的位元線、未被選單元的位元線、被選單元的源極線及未被選單元的源極線上之電壓。列表示讀取、抹除及程式化的操作。

表5：圖10的VMM陣列1000之操作

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

【0050】 圖11描繪神經元VMM陣列1100，其特別適用於圖2所示之記憶體單元210，並且用作輸入層與下一層之間的突觸及神經元的部件。VMM陣列1100包括非揮發性記憶體單元的記憶體陣列1103、第一非揮發性參考記憶體單元的參考陣列1101及第二非揮發性參考記憶體單元的參考陣列1102。參考陣列1101及1102在VMM陣列1100的列方向上延伸。除在VMM陣列1100中，字元線在垂直方向上延伸外，VMM

陣列與VMM 1000相似。這裡，在字元線(WLA0、WLB0、WLA1、WLB1、WLA2、WLB2、WLA3、WLB3)上提供輸入，並且在讀取操作期間輸出出現在源極線(SL0，SL1)上。在每條源極線上之電流執行來自連接至那條特定源極線之記憶體單元的所有電流之求和函數。

【0051】 表6描繪VMM陣列1100的操作電壓及電流。表中之行表示在被選單元的字元線、未被選單元的字元線、被選單元的位元線、未被選單元的位元線、被選單元的源極線及未被選單元的源極線上之電壓。列表示讀取、抹除及程式化的操作。

表6：圖11的VMM陣列1100之操作

	WL	WL-unsel	BL	BL-unsel	SL	SL-unsel
讀取	1-3.5 V	-0.5 V/0 V	0.6-2 V	0.6 V-2 V/0 V	~0.3-1 V(I神經元)	0 V
抹除	~5-13 V	0 V	0 V	0 V	0 V	禁止 SL(~4-8 V)
程式化	1-2 V	-0.5 V/0 V	0.1-3 uA	Vinh~2.5 V	4-10 V	0-1 V/FLT

【0052】 圖12描繪神經元VMM陣列1200，其特別適用於圖3所示之記憶體單元310，並且用作輸入層與下一層之間的突觸及神經元的部件。VMM陣列1200包括非揮發性記憶體單元的記憶體陣列1203、第一非揮發性參考記憶體單元的參考陣列1201及第二非揮發性參考記憶體單元的參考陣列1202。參考陣列1201及1202用於將流入端子BLR0、BLR1、BLR2及BLR3的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係經由多工器1212(部分被顯示)與經由BLR0、BLR1、BLR2及BLR3流入它們的電流輸入以二極體形式連接。多工器1212各自包括個別多工器1205及疊接電晶體1204，以在讀取操作期間確保第一及第二非揮發性參考記憶體單元

中之每一者的位元線(諸如BLR0)上的固定電壓。將參考單元調整至目標參考位準。

【0053】 記憶體陣列1203提供兩個用途。第一，它儲存將被VMM陣列1200使用的權重。第二，記憶體陣列1203有效地將輸入(被提供至端子BLR0、BLR1、BLR2及BLR3的電流輸入；參考陣列1201及1202將這些電流輸入轉換成輸入電壓，以供應至控制閘極(CG0、CG1、CG2及CG3))乘以記憶體陣列中所儲存之權重，然後將所有結果(單元電流)相加，以產生輸出，所述輸出出現在BL0-BLN且將是下一層的輸入或最後一層的輸入。藉由執行乘法及加法函數，記憶體陣列不需要個別的乘法及加法邏輯電路，並且還具有功率效率。這裡，輸入被提供在控制閘極線(CG0、CG1、CG2及CG3)上，並且輸出在讀取操作期間出現在位元線(BL0-BLN)上。在每條位元線上之電流執行來自連接至那條特定位元線之記憶體單元的所有電流之求和函數。

【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整。亦即，抹除及然後部分程式化每個非揮發性記憶體單元，直到達到浮動閘極上的期望電荷為止。如果使太多電荷置於浮動閘極上(使得錯誤值儲存在單元中)，則抹除單元並且重新開始部分程式化操作的順序。如圖所示，共享同一個抹除閘極(例如，EG0或EG1)的兩列要一起被抹除(亦稱為頁抹除)，之後，部分程式化每個單元，直到達到浮動閘極上之期望電荷為止。

【0055】 表7描繪VMM陣列1200的操作電壓及電流。表中之行表示在被選單元的字元線、未被選單元的字元線、被選單元的位元線、

未被選單元的位元線、被選單元的控制閘極、與被選單元相同的區段中之未被選單元的控制閘極、與被選單元不同的區段中之未被選單元的控制閘極、被選單元的抹除閘極、未被選單元的抹除閘極、被選單元的源極線及未被選單元的源極線上之電壓。列表表示讀取、抹除及程式化的操作。

表7：圖12的VMM陣列1200之操作

	WL	WL-unsel	BL	BL-unsel	CG	在同一個區段中之CG-unsel	CG-unsel	EG	EG-unsel	SL	SL-unsel
讀取	1.0-2V	-0.5V/0V	0.6-2V (I神經元)	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描繪神經元VMM陣列1300，其特別適用於圖3所示之記憶體單元310，並且用作輸入層與下一層之間的突觸及神經元的部件。VMM陣列1300包括非揮發性記憶體單元的記憶體陣列1303、第一非揮發性參考記憶體單元的參考陣列1301及第二非揮發性參考記憶體單元的參考陣列1302。EG線EGR0、EG0、EG1及EGR1垂直延伸，而CG線CG0、CG1、CG2及CG3以及SL線WL0、WL1、WL2及WL3水平延伸。除VMM陣列1300實施雙向調整外，VMM陣列1300與VMM陣列1200相似，其中每個個別單元可以完全被抹除、部分被程式化及根據需要部分被抹除，以因個別EG線的使用而在浮動閘極上達到所需的電荷量。如圖所示，參考陣列1301及1302將端子BLR0、BLR1、BLR2及BLR3中之輸入電流轉換成要施加至列方向上的記憶體單元之控制閘極電壓CG0、CG1、CG2及CG3(藉由經由多工器1314以二極體形式連接之參考

單元的作用)。電流輸出(神經元)位於位元線BL0-BLN中，其中每條位元線計算來自與那條特定位元線連接之非揮發性記憶體單元的所有電流之總和。

【0057】 表8描繪VMM陣列1300的操作電壓及電流。表中之行表示在被選單元的字元線、未被選單元的字元線、被選單元的位元線、未被選單元的位元線、被選單元的控制閘極、與被選單元相同的區段中之未被選單元的控制閘極、與被選單元不同的區段中之未被選單元的控制閘極、被選單元的抹除閘極、未被選單元的抹除閘極、被選單元的源極線及未被選單元的源極線上之電壓。列表示讀取、抹除及程式化的操作。

表8：圖13的VMM陣列1300之操作

	WL	WL-unsel	BL	BL-unsel	CG	同一個區段中之CG-unsel	CG-unsel	EG	EG-unsel	SL	SL-unsel
讀取	1.0-2V	-0.5V/0V	0.6-2V (I神經元)	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	-0.5V/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

【0058】 圖22描繪神經元VMM陣列2200，其特別適合於圖2所示之記憶體單元210，並且用以作為輸入層與下一層之間的突觸及神經元的部件。在VMM陣列2200中，在位元線BL₀、…、BL_N上分別接收輸入INPUT₀、…、INPUT_N，並且在源極線SL₀、SL₁、SL₂及SL₃上分別產生輸出OUTPUT₁、OUTPUT₂、OUTPUT₃及OUTPUT₄。

【0059】 圖23描繪神經元VMM陣列2300，其特別適合於圖2所示之記憶體單元210，並且用以作為輸入層與下一層之間的突觸及神經

元的部件。在此實例中，在源極線 SL_0 、 SL_1 、 SL_2 及 SL_3 上分別接收輸入 $INPUT_0$ 、 $INPUT_1$ 、 $INPUT_2$ 及 $INPUT_3$ ，並且在位元線 BL_0 、 $\dots$ 、 BL_N 上產生輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 。

【0060】 圖24描繪神經元VMM陣列2400，其特別適合於圖2所示之記憶體單元210，並且用以作為輸入層與下一層之間的突觸及神經元的部件。在此實例中，在字元線 WL_0 、 $\dots$ 、 WL_M 上分別接收輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ ，並且在位元線 BL_0 、 $\dots$ 、 BL_N 上產生輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 。

【0061】 圖25描繪神經元VMM陣列2500，其特別適合於圖3所示之記憶體單元310，並且用以作為輸入層與下一層之間的突觸及神經元的部件。在此實例中，在字元線 WL_0 、 $\dots$ 、 WL_M 上分別接收輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_M$ ，並且在位元線 BL_0 、 $\dots$ 、 BL_N 上產生輸出 $OUTPUT_0$ 、 $\dots$ 、 $OUTPUT_N$ 。

【0062】 圖26描繪神經元VMM陣列2600，其特別適合於圖4所示之記憶體單元410，並且用以作為輸入層與下一層之間的突觸及神經元的部件。在此實例中，在垂直控制閘極線 CG_0 、 $\dots$ 、 CG_N 上分別接收輸入 $INPUT_0$ 、 $\dots$ 、 $INPUT_n$ ，並且在源極線 SL_0 及 SL_1 上產生輸出 $OUTPUT_1$ 及 $OUTPUT_2$ 。

【0063】 圖27描繪神經元VMM陣列2700，其特別適合於圖4所示之記憶體單元410，並且用以作為輸入層與下一層之間的突觸及神經元的部件。在此實例中，在分別耦接至位元線 BL_0 、 $\dots$ 、 BL_N 之位元線控制閘2701-1、2701-2、 $\dots$ 、2701-(N-1)及2701-N的閘極上分別接收輸入

$INPUT_0, \dots, INPUT_N$ 。在源極線 SL_0 及 SL_1 上產生示例性輸出 $OUTPUT_1$ 及 $OUTPUT_2$ 。

【0064】 圖28描繪神經元VMM陣列2800，其特別適合於圖3所示之記憶體單元310、圖5所示之記憶體單元510及圖7所示之記憶體單元710，並且用以作為輸入層與下一層之間的突觸及神經元的部件。在此實例中，在字元線 $WL_0, \dots, WL_M$ 上接收輸入 $INPUT_0, \dots, INPUT_M$ ，並且在位元線 $BL_0, \dots, BL_N$ 上分別產生輸出 $OUTPUT_0, \dots, OUTPUT_N$ 。

【0065】 圖29描繪神經元VMM陣列2900，其特別適合於圖3所示之記憶體單元310、圖5所示之記憶體單元510及圖7所示之記憶體單元710，並且用以作為輸入層與下一層之間的突觸及神經元的部件。在此實例中，在控制閘極線 $CG_0, \dots, CG_M$ 上接收輸入 $INPUT_0, \dots, INPUT_M$ 。在垂直源極線 $SL_0, \dots, SL_N$ 上分別產生輸出 $OUTPUT_0, \dots, OUTPUT_N$ ，其中每條源極線 SL_i 耦接至第 i 行中之所有記憶體單元的源極線。

【0066】 圖30描繪神經元VMM陣列3000，其特別適合於圖3所示之記憶體單元310、圖5所示之記憶體單元510及圖7所示之記憶體單元710，並且用以作為輸入層與下一層之間的突觸及神經元的部件。在此實例中，在控制閘極線 $CG_0, \dots, CG_M$ 上接收輸入 $INPUT_0, \dots, INPUT_M$ 。在垂直位元線 $BL_0, \dots, BL_N$ 上分別產生輸出 $OUTPUT_0, \dots, OUTPUT_N$ ，其中每條位元線 BL_i 耦接至第 i 行中之所有記憶體單元的位元線。

長短期記憶體

【0067】 習知技藝包括稱為長短期記憶體(LSTM)的概念。LSTM單元通常用於神經網路中。LSTM允許神經網路在預定任意時間

間隔內記住資訊，並在後續操作中使用那個資訊。傳統的LSTM單元包括單元、輸入閘極、輸出閘極及遺忘閘極。三個閘極調整進出單元的資訊流及在LSTM中記住資訊的時間間隔。VMM在LSTM單元中係特別有用的。

【0068】 圖14描繪示例性LSTM 1400。此實例中的LSTM 1400包括單元1401、1402、1403及1404。單元1401接收輸入向量 x_0 ，並產生輸出向量 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的輸出向量(隱藏狀態) h_0 及單元狀態 c_0 ，並產生輸出向量 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的輸出向量(隱藏狀態) h_1 及單元狀態 c_1 ，並產生輸出向量 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的輸出向量(隱藏狀態) h_2 及單元狀態 c_2 ，並產生輸出向量 h_3 。可以使用額外的單元，並且具有四個單元的LSTM僅是一個實例。

【0069】 圖15描繪LSTM單元1500的示例性實施，其可以用於圖14中之單元1401、1402、1403及1404。LSTM單元1500接收輸入向量 $x(t)$ 、來自前一個單元之單元狀態向量 $c(t-1)$ 及來自前一個單元之輸出向量 $h(t-1)$ ，並產生單元狀態向量 $c(t)$ 及輸出向量 $h(t)$ 。

【0070】 LSTM單元1500包括sigmoid函數裝置1501、1502及1503，每個sigmoid函數裝置應用0與1之間的數字，以控制輸入向量中之每個分量有多少被允許直至輸出向量。LSTM單元1500亦包括用以將雙曲正切函數應用於輸入向量的tanh裝置1504及1505、用以將兩個向量相乘的乘法裝置1506、1507及1508以及用以將兩個向量相加的加法裝置

1509。輸出向量 $h(t)$ 可以提供給系統中的下一個LSTM單元，或者亦可以出於其它目的對其進行存取。

【0071】 圖16描繪LSTM單元1600，其為LSTM單元1500的實施之一個實例。為方便讀者，在LSTM單元1600中使用與LSTM單元1500相同的編號。sigmoid函數裝置1501、1502及1503以及tanh裝置1504各自包括多個VMM陣列1601及激勵函數區塊1602。因此，可以看出VMM陣列在某些神經網路系統中使用之LSTM單元中係特別有用的。乘法裝置1506、1507及1508以及加法裝置1509以數位方式或以類比方式來實施。激勵函數區塊1602可以數位方式或類比方式來實施。

【0072】 圖17顯示LSTM單元1600的一個替代方案(以及LSTM單元1500實施的另一個實例)。在圖17中，Sigmoid函數裝置1501、1502及1503以及tanh裝置1504以時間多工方式共享同一個實體硬體(VMM陣列1701及激勵函數區塊1702)。LSTM單元1700亦包括：乘法裝置1703，用於將兩個向量相乘；加法裝置1708，用於將兩個向量相加；tanh裝置1505(其包括激勵函數區塊1702)；暫存器1707，其在 $i(t)$ 從sigmoid函數區塊1702輸出時儲存數值 $i(t)$ ；暫存器1704，其在數值 $f(t)*c(t-1)$ 從乘法裝置1703經由多工器1710輸出時儲存該數值；暫存器1705，其在數值 $i(t)*u(t)$ 從乘法裝置1703經由多工器1710輸出時儲存該數值；及暫存器1706，其在數值 $o(t)*\tilde{c}(t)$ 從乘法裝置1703經由多工器1710輸出時儲存該數值；以及多工器1709。

【0073】 LSTM單元1600包含多組VMM陣列1601及個別的激勵函數區塊1602，而LSTM單元1700僅包含一組VMM陣列1701及激勵

函數區塊1702，它們在LSTM單元1700的具體例中用於表示多層。LSTM單元1700將需要比LSTM單元1600少的空間，因為相較於LSTM單元1600，LSTM單元1700只需要1/4空間用於VMM及激勵函數區塊。

【0074】 可以進一步理解，LSTM單元通常將包括多個VMM陣列，每個VMM陣列需要由VMM陣列外部的某些電路區塊(例如，加法器及激勵函數區塊以及高電壓產生區塊)提供的功能。對每個VMM陣列提供個別的電路區塊，將在半導體裝置內需要大量的空間，並且會有些沒有效率。因此，下面所描述的實例減少VMM陣列本身外部所需的電路。

閘控遞歸單元

【0075】 可以將類比VMM實施用於GRU(閘控遞歸單元)系統。GRU係遞歸神經網路中的閘控機制。除GRU單元通常包含比LSTM單元少的組件外，GRU與LSTM相似。

【0076】 圖18描繪示例性GRU 1800。此實例中之GRU 1800包括單元1801、1802、1803及1804。單元1801接收輸入向量 x_0 ，並產生輸出向量 h_0 。單元1802接收輸入向量 x_1 及來自單元1801的輸出向量 h_0 ，並產生輸出向量 h_1 。單元1803接收輸入向量 x_2 及來自單元1802的輸出向量(隱藏狀態) h_1 ，並產生輸出向量 h_2 。單元1804接收輸入向量 x_3 及來自單元1803的輸出向量(隱藏狀態) h_2 ，並產生輸出向量 h_3 。可以使用額外的單元，並且具有四個單元的GRU僅是一個實例。

【0077】 圖19描繪GRU單元1900的示例性實施，其可以用於圖18之單元1801、1802、1803及1804。GRU單元1900接收輸入向量 $x(t)$ 及來自前一個GRU單元之輸出向量 $h(t-1)$ ，並產生輸出向量 $h(t)$ 。GRU單元

1900包括sigmoid函數裝置1901及1902，每個sigmoid函數裝置應用0與1之間的數字於來自輸出向量 $h(t-1)$ 及輸入向量 $x(t)$ 的分量。GRU單元1900亦包括用以將雙曲正切函數應用於輸入向量的tanh裝置1903、用以將兩個向量相乘的複數個乘法裝置1904、1905及1906、用以將兩個向量相加的加法裝置1907以及用以從1減去輸入來產生輸出的互補裝置1908。

【0078】 圖20描繪GRU單元2000，其是GRU單元1900的實施之一個實例。為方便讀者，在GRU單元2000中使用與GRU單元1900相同的編號。從圖20可以看出，sigmoid函數裝置1901及1902以及tanh裝置1903各自包括多個VMM陣列2001及激勵函數區塊2002。因此，可以看出VMM陣列特別用於某些神經網路系統中使用之GRU單元中。乘法裝置1904、1905及1906、加法裝置1907以及互補裝置1908以數位方式或以類比方式來實施。激勵函數區塊2002可以數位方式或類比方式來實施。

【0079】 圖21顯示GRU單元2000的一個替代方案(以及GRU單元1900實施的另一個實例)。在圖21中，GRU單元2100利用VMM陣列2101及激勵函數區塊2102，激勵函數區塊2102在配置成為Sigmoid函數時應用0與1之間的數字，以控制輸入向量中之每個分量有多少被允許直至輸出向量。在圖21中，Sigmoid函數裝置1901及1902以及tanh裝置1903以時間多工方式共享同一個實體硬體(VMM陣列2101及激勵函數區塊2102)。GRU單元2100亦包括：乘法裝置2103，用於將兩個向量相乘；加法裝置2105，用於將兩個向量相加；互補裝置2109，用於從1減去

輸入，以產生輸出；多工器2104；暫存器2106，用於當數值 $h(t-1)*r(t)$ 從乘法裝置2103經由多工器2104輸出時，保持該數值；暫存器2107，用於當數值 $h(t-1)*z(t)$ 從乘法裝置2103經由多工器2104輸出時，保持該數值；以及暫存器2108，用於當數值 $h^{\wedge}(t)*(1-z(t))$ 從乘法裝置2103經由多工器2104輸出時，保持該數值。

【0080】 GRU單元2000包含多組VMM陣列2001及激勵函數區塊2002，而GRU單元2100僅包含一組VMM陣列2101及激勵函數區塊2102，它們在GRU單元2100的具體例中用於表示多層。GRU單元2100將需要比GRU單元2000少的空間，因為相較於GRU單元2000，GRU單元2100只需要1/3空間用於VMM及激勵函數區塊。

【0081】 可以進一步理解，GRU系統通常將包括多個VMM陣列，每個VMM陣列需要由VMM陣列外部的某些電路區塊(例如，加法器及激勵函數區塊以及高電壓產生區塊)提供的功能。對每個VMM陣列提供個別的電路區塊，將在半導體裝置內需要大量的空間，並且會有點沒效率。因此，下面所描述的具體例減少VMM陣列本身外部所需的電路。

【0082】 VMM陣列的輸入可以是類比位準、二進制位準、脈衝、時間調變脈衝或數位位元(在這種情況下，需要DAC將數位位元轉換為適當的輸入類比位準)，以及輸出可以是類比位準、二進制位準、時序脈衝、脈衝或數位位元(在這種情況下，需要輸出ADC將輸出類比位準轉換為數位位元)。

【0083】 通常，對於一個VMM陣列中之每個記憶體單元，每個權重W可以由單一記憶體單元或由一個差分單元或由兩個混合記憶體單

元(平均2個單元)來實施。在差分單元的情況下，需要兩個記憶體單元來實施權重 W 成為差分權重($W=W+-W-$)。在兩個混合記憶體單元方面，需要兩個記憶體單元來實施權重 W 成為兩個單元的平均值。

【0084】 圖31描繪VMM系統3100。在一些實例中，將儲存在VMM陣列中的權重 W 儲存為差分對 $W+$ (正權重)及 $W-$ (負權重)，其中 $W=(W+)-(W-)$ 。在VMM系統3100中，一半的位元線稱為 $W+$ 線，即連接至將儲存正權重 $W+$ 之記憶體單元的位元線，而另一半的位元線稱為 $W-$ 線，即連接至實現負權重 $W-$ 之記憶體單元的位元線。 $W-$ 線以交替方式散置在 $W+$ 線之間。減法運算由從 $W+$ 線及 $W-$ 線接收電流之求和電路(例如，求和電路3101及3102)來執行。 $W+$ 線的輸出與 $W-$ 線的輸出組合在一起以有效地為所有($W+$ 、 $W-$)線對的每對($W+$ 、 $W-$)單元提供 $W=W+-W-$ 。雖然以上已經針對 $W-$ 線以交替方式散置在 $W+$ 線之間來進行描述，但在其它實例中 $W+$ 線及 $W-$ 線可以任意地位於陣列中的任何位置。

【0085】 圖32描繪另一個實例。在VMM系統3210中，正權重 $W+$ 在第一陣列3211中實現，負權重 $W-$ 在第二陣列3212中實現，第二陣列3212與第一陣列分開，並且所得權重由求和電路3213適當地組合在一起。

【0086】 圖33描繪VMM系統3300。將儲存在VMM陣列中的權重 W 儲存為差分對 $W+$ (正權重)及 $W-$ (負權重)，其中 $W=(W+)-(W-)$ 。VMM系統3300包括陣列3301及陣列3302。陣列3301及3302中之每個陣列的一半位元線稱為 $W+$ 線，即連接至將儲存正權重 $W+$ 之記憶體單元的位元線，而陣列3301及3302中之每個陣列的另一半位元線稱為 $W-$ 線，即

連接至實現負權重 W^- 之記憶體單元的位元線。 W^- 線以交替方式散置在 W^+ 線之間。減法運算由從 W^+ 線及 W^- 線接收電流的求和電路(例如, 求和電路3303、3304、3305及3306)來執行。來自每個陣列3301、3302之 W^+ 線的輸出與 W^- 線的輸出分別組合在一起, 以有效地為所有(W^+ 、 W^-)線對的每對(W^+ 、 W^-)單元提供 $W=W^+-W^-$ 。此外, 來自每個陣列3301及3302的 W 值可以由求和電路3307及3308來進一步組合, 使得每個 W 值是來自陣列3301的 W 值減去來自陣列3302的 W 值之結果, 這意味著求和電路3307及3308的最終結果是兩個微分值的微分值。

【0087】 類比神經記憶體系統中使用之每個非揮發性記憶體單元都將被抹除及程式化, 以在浮動閘極中保持非常特定及精確的電荷量, 亦即, 電子數量。例如, 每個浮動閘極應該保存 N 個不同值中之一, 其中 N 是由每個單元可以指示的不同權重的數量。 N 的實例包括16、32、64、128及256。

【0088】 圖34描繪習知技藝的VMM系統3400, 其包括兩個不同的陣列區塊(類比陣列區塊3405及數位陣列區塊3415)。類比陣列區塊3405包括類比陣列3401、列解碼器3402及高電壓解碼器3403。類比陣列3401儲存如上所述之 N 個不同類比值中之一。數位陣列區塊3415包括數位陣列3411、列解碼器3412及高電壓解碼器3413。數位陣列區塊3415可用於儲存系統資料、組態資料、作業系統資料(OS)以及VMM系統3400操作所需的其它資料。例如, 數位陣列3411可以儲存數位資料, 例如, 用戶ID、微調位元、製造資訊、安全代碼、OS代碼及VMM系統3400使

用的其它資訊。在另一個實例中，VMM系統3400可以包括一個單獨的數位非揮發性記憶體巨集或外部晶片。

【0089】 因為類比陣列3401儲存類比值(多級值)，所以相較於數位陣列3411的讀取操作(其中每個單元儲存「1」或「0」)，在讀取操作期間將準確的電流範圍(歸因於在期望的低功率下之多級)施加至位元線。讀取操作期間使用之電流位準的差異會導致不良後果，例如，來自數位單元的高電流單元之洩漏(這是高速讀取所需要的，會影響讀取操作的準確性)、偏壓條件(對於類比位準與數位位準來說是不同的)及其它後果。

【0090】 所需要的是一種用於在人工神經網路系統中提供類比陣列及數位陣列並同時減少洩漏及其它不良後果的改良架構。

【發明內容】

【0091】 描述用於提供包含類比陣列及數位陣列之人工神經網路系統的許多實例。

【圖式簡單說明】

【0092】 圖1係繪示人工神經網路之示圖。

【0093】 圖2描繪習知技藝的分離式閘極快閃記憶體單元。

【0094】 圖3描繪另一個習知技藝的分離式閘極快閃記憶體單元。

【0095】 圖4描繪另一個習知技藝的分離式閘極快閃記憶體單元。

【0096】 圖5描繪另一個習知技藝的分離式閘極快閃記憶體單元。

【0097】 圖6係繪示利用一個或多個非揮發性記憶體陣列之示例性人工神經網路的不同層級之示圖。

【0098】 圖7係繪示VMM系統的方塊圖。

【0099】 圖8係繪示使用一個或多個VMM系統之示例性人工神經網路的方塊圖。

【0100】 圖9描繪VMM系統的另一個實例。

【0101】 圖10描繪VMM系統的另一個實例。

【0102】 圖11描繪VMM系統的另一個實例。

【0103】 圖12描繪VMM系統的另一個實例。

【0104】 圖13描繪VMM系統的另一個實例。

【0105】 圖14描繪習知技藝的長短期記憶體系統。

【0106】 圖15描繪用於長短期記憶體系統中之一個示例性單元。

【0107】 圖16描繪圖15的單元之一個示例性實施。

【0108】 圖17描繪圖15的單元之另一個示例性實施。

【0109】 圖18描繪習知技藝的閘控遞歸單元系統(gated recurrent unit system)。

【0110】 圖19描繪用於閘控遞歸單元系統中之一個示例性單元。

【0111】 圖20描繪圖19的單元之一個示例性實施。

- 【0112】 圖21描繪圖19的單元之另一個示例性實施。
- 【0113】 圖22描繪VMM系統的另一個實例。
- 【0114】 圖23描繪VMM系統的另一個實例。
- 【0115】 圖24描繪VMM系統的另一個實例。
- 【0116】 圖25描繪VMM系統的另一個實例。
- 【0117】 圖26描繪VMM系統的另一個實例。
- 【0118】 圖27描繪VMM系統的另一個實例。
- 【0119】 圖28描繪VMM系統的另一個實例。
- 【0120】 圖29描繪VMM系統的另一個實例。
- 【0121】 圖30描繪VMM系統的另一個實例。
- 【0122】 圖31描繪VMM系統的另一個實例。
- 【0123】 圖32描繪VMM系統的另一個實例。
- 【0124】 圖33描繪VMM系統的另一個實例。
- 【0125】 圖34描繪包括數位陣列之VMM系統的一個實例。
- 【0126】 圖35描繪VMM系統的一個實例。
- 【0127】 圖36描繪包括數位陣列之VMM系統的一個實例。
- 【0128】 圖37描繪包括數位陣列之VMM系統的一個實例。
- 【0129】 圖38描繪圖37的VMM系統之額外態樣。
- 【0130】 圖39A、39B及39C描繪圖37的VMM系統之示例性設計。
- 【0131】 圖40描繪包括數位陣列的VMM系統之一個實例。
- 【0132】 圖41描繪數位陣列的一個實例。

【0133】 圖42A描繪差分電流至電壓轉換器。

【0134】 圖42B描繪差分連續位址暫存器類比至數位轉換器。

【0135】 圖43描繪類比陣列及數位陣列的一個示例性讀取操作。

【實施方式】

【0136】 本文所述的人工神經網路利用CMOS技術與非揮發性記憶體陣列的組合。

VMM系統概述

【0137】 圖35描繪VMM系統3500的方塊圖。VMM系統3500包括VMM陣列3501(其包括類比記憶體單元及數位記憶體單元)、列解碼器3502、高電壓解碼器3503、行解碼器3504、位元線驅動器3505、輸入電路3506、輸出電路3507、控制邏輯3508及偏壓產生器3509。VMM系統3500進一步包括高電壓產生區塊3510，其包括電荷泵3511、電荷泵調節器3512及高電壓類比精確位準產生器3513。VMM系統3500進一步包括(程式化/抹除，或權重調整)演算法控制器3514、類比電路3515、控制引擎3516(其可執行諸如算術函數、激勵函數、嵌入式微控制器邏輯的功能，但不限於此)及測試控制邏輯3517。下文描述的系統及方法可在VMM系統3500中實現。

【0138】 輸入電路3506可以包括諸如DAC(數位至類比轉換器)、DPC(數位至脈衝轉換器，數位至時間調變脈衝轉換器)、AAC(類比至類比轉換器，例如，電流至電壓轉換器、對數轉換器)、PAC(脈衝至類比位準轉換器)或任何其它類型的轉換器之電路。輸入電路3506可以實

施正規化、線性或非線性縮放函數及算術函數中的一個或多個。輸入電路3506可以對輸入位準實施溫度補償函數。輸入電路3506可以實施諸如ReLU或sigmoid的激勵函數。輸出電路3507可以包括諸如ADC(類比至數位轉換器，以將神經元類比輸出轉換成數位位元)、AAC(類比至類比轉換器，例如，電流至電壓轉換器、對數轉換器)、APC(類比至脈衝轉換器、類比至時間調變脈衝轉換器)或任何其它類型的轉換器之電路。輸出電路3507可以實施諸如整流線性激勵函數(ReLU)或sigmoid的激勵函數。輸出電路3507可以對神經元輸出實施統計正規化、正則化、縮放/增益函數、統計修整及算術函數(例如，加、減、除、乘、移位、對數)中的一個或多個。輸出電路3507可以對神經元輸出或陣列輸出(例如，位元線輸出)實施溫度補償函數，以便例如藉由保持IV斜率大致相同來保持陣列的功率消耗近似恆定或提高陣列(神經元)輸出的精度。

【0139】 如上所述，神經網路可能包括許多不同的層，並且在每一層內，可以執行許多計算，這些計算涉及該層內一個或多個陣列中儲存的權重值。一些層將比其它層使用得更多，並且可以理解，根據它們的高使用頻率，這樣的層對神經網路的整體精度更為重要。

【0140】 圖36描繪VMM系統3600，其包括類比陣列3601、列解碼器3602、高電壓解碼器3603、數位陣列3604、列解碼器3605及高電壓解碼器3606。類比陣列3601及數位陣列3604共享相同的擴散及金屬互連位元線。與習知技藝不同，數位陣列3604具有其自己的列解碼器(列解碼器3605)及其自己的高電壓解碼器(3606)，這允許同時施加第一組電壓和

第二組電壓分別至類比陣列3601及數位陣列3604。可選地，類比陣列3601及數位陣列3604被製造在同一個半導體晶粒上。

【0141】 數位陣列3604可以可選地包括第一陣列(可稱為使用者資料陣列3607)，其包括用於儲存由外部源儲存及檢索的數位資料(例如，使用者資料或用於電腦的作業系統代碼)之列；以及第二陣列(可稱為系統資料陣列3608)，其包括用於儲存數位系統資料(亦即，由VMM系統3600本身使用且不被外部源儲存及檢索的資料)之列，其有時稱為數位非揮發性暫存器(NVR)或資訊列或資訊陣列。可以儲存在系統資料陣列3608中之資料的實例包括使用者ID、微調位元、組態位元、製造資訊、安全代碼、密碼、鎖定位元及其它系統資料。與使用者資料陣列3607相比，可以放寬系統資料陣列3608的效能。

【0142】 例如，系統資料陣列3608的內容可以在供電時或在系統操作開始時被讀取一次，之後在操作期間不被讀取。作為另一個實例，系統資料陣列3608的內容在其整個使用期限可能僅被抹除或程式化幾次。因此，系統資料陣列3608的讀取、程式化或抹除操作可以以比使用者資料陣列慢的速度發生，而沒有太多效能損失，因為系統資料陣列很少使用。這種速度差異可以在VMM系統3600中實現，例如，以在系統資料陣列3608上操作時節省電力。

【0143】 作為另一個實例，非揮發性記憶體單元具有稱為耐久性的特性，其意指非揮發性記憶體單元在退化到不再可靠或不再可用的時間點之前可以被程式化或抹除的次數。因此，系統資料陣列3608可以具有比使用者資料陣列3607的非揮發性記憶體單元耐久性低的非揮發

性記憶體單元，因為系統資料陣列3608在操作期間將比使用者資料陣列3607使用得少得多。

【0144】 表9描繪VMM系統3600中使用的示例性操作電壓，其中CG-main、EG-main、BL-main及SL-main係由列解碼器3602及高電壓解碼器3603分別施加至類比陣列3601中之一個或多個記憶體單元的控制閘極端子、抹除閘極端子、位元線端子及源極線端子的第一組電壓，以及CG-DIG、EG-DIG、BL-DIG及SL-DIG係由列解碼器3605及高電壓解碼器3606分別施加至數位陣列3604中之一個或多個單元的控制閘極端子、抹除閘極端子、位元線端子及源極線端子的第二組電壓。

表9：圖36中之VMM系統3600的操作電壓

動作	CG-main	EG-main	BL-main	SL-main	CG-DIG	EG-DIG	BL-DIG	SL-DIG
神經讀取	1.5V	0V	0.6V	$\geq 0V$	0V	0V	0.6V	$\geq 0V$
數位讀取	0V	0V	0.9V	0V	1.5V-2.5V	0V-1.5V	0.9V	0V

【0145】 將不同組的電壓施加至類比陣列3601的記憶體單元及數位陣列3604的記憶體單元之能力增強類比陣列3601的神經讀取操作及數位陣列3604的數位讀取操作期間的效能，例如，減少神經讀取操作期間的洩漏(由關閉與類比單元共享相同位元線之數位單元的高電流之能力引起的)以及由於更高的電流位準所造成之數位單元的更高速度。

【0146】 圖37描繪示例性VMM系統3700，其包括類比陣列3701、列解碼器3702、高電壓解碼器3703、數位陣列3704(可選地包括使用者資料陣列3707及系統資料陣列3708)、列解碼器3705及高電壓解碼器3706。VMM系統3700相似於VMM系統3600，除耦接至類比陣列

3701的位元線與耦接至數位陣列3704的位元線斷開外，這意味著類比陣列3701與數位陣列3704具有個別的位元線。然而，類比陣列3701的位元線與數位陣列3704的位元線共享相同的擴散層。在另一個實例中，類比陣列3701的位元線及數位陣列3704的位元線除了斷開之外還具有不同的擴散層。擴散可以在類比位元線與數位位元線之間用一個或多個虛擬列斷開。可選地，類比陣列3701及數位陣列3704被製造在同一半導體晶粒上。

【0147】 表10描繪VMM系統3700中使用的操作電壓，其中CG-main、EG-main、BL-main及SL-main係由列解碼器3702及高電壓解碼器3703分別施加至類比陣列3701中之一個或多個記憶體單元的控制閘極端子、抹除閘極端子、位元線端子及源極線端子的第一組電壓，以及CG-IFR、EG-IFR、BL-IFR及SL-IFR係由列解碼器3705及高電壓解碼器3706分別施加至數位陣列3704中之一個或多個記憶體單元的控制閘極端子、抹除閘極端子、位元線端子及源極線端子的第二組電壓。

表10：圖37中之VMM系統3700的操作電壓

動作	CG-main	EG-main	BL-main	SL-main	CG-DIG	EG-DIG	BL-DIG	SL-DIG
神經讀取	1.5V	0V	0.6V	>=0V	0V	0V	0V	0V
數位讀取	0V	0V	0V	0V	2.5V	1.5V	0.9V	0V

【0148】 圖43描繪類比陣列(例如，圖36及37中的類比陣列3601及3701)及數位陣列(例如，圖36及37中的數位陣列3604及3704)的示例性讀取操作4300。在操作4301中，在類比陣列的讀取期間，第一列解碼器及第一電壓列解碼器將第一組電壓施加至類比陣列。在操作4302

中，在數位陣列的讀取期間，第二列解碼器及第二電壓列解碼器將第二組電壓施加至數位陣列，其中第二組電壓不同於第一組電壓。

【0149】 將不同組電壓施加至類比陣列3701中之記憶體單元及數位陣列3704中之記憶體單元的能力增強類比陣列3701的神經讀取操作及數位陣列3704的數位讀取操作期間的效能，例如，減少神經讀取操作期間的洩漏(由關閉與類比單元共享相同位元線之數位單元的高電流之能力引起的)以及由於更高的電流位準所造成之數位單元的更高速度。

【0150】 圖38描繪VMM系統3700的附加態樣，其進一步包括行解碼器3801及行解碼器3802。行解碼器3801耦接至與類比陣列3701的行耦接之位元線BL0、...、BLn。行解碼器3802耦接至與數位陣列3704的行耦接之位元線BLD0、...、BLDn。數位陣列3704可選地包括使用者資料陣列3707及系統資料陣列3708。因此，類比陣列3701及數位陣列3704具有個別的行解碼器，進一步減少神經讀取操作期間的洩漏。

【0151】 圖39A及39B分別描繪VMM系統3901、3902，它們是VMM系統3700的示例性設計，以及圖39C描繪VMM系統3903，它是VMM系統3600的示例性設計。

【0152】 在圖39A中，VMM系統3901包括類比陣列3701及數位陣列3704。示例性位元線BL0耦接至類比陣列3701，而示例性位元線BLIFR0耦接至數位陣列3704。耦接至類比陣列3701的位元線BL0可以稱為類比陣列位元線BL0，使用類比陣列3701中的金屬層M1至M4，金屬層M2至M4亦連接至週邊電路(例如，行解碼器)，而耦接至數位陣列3704的位元線BLIFR0可以稱為數位陣列位元線BLIFR0，僅使用金屬層

M1。類比陣列3701及數位陣列3704係同一物理陣列3904的一部分。物理陣列3904包括基板3905及擴散層3906，它們均由類比陣列3701及數位陣列3704共享。物理陣列3904藉由使用用於類比陣列3701及數位陣列3704之斷開的金屬內連線來分割。例如，可以看到，類比陣列位元線BL0與數位陣列位元線BLIFR0彼此斷開。對於類比陣列3701及數位陣列3704使用相同的物理陣列3904減少在使用分開的物理陣列時會發生之製程一致性及面積負擔的任何物理效應。

【0153】 在圖39B中，VMM系統3902包括類比陣列3701及數位陣列3704。示例性位元線BL0耦接至類比陣列3701，並且可以稱為類比陣列位元線BL0，而示例性位元線BLIFR0耦接至數位陣列3704，並且可以稱為數位陣列位元線BLIFR0。類比陣列位元線BL0使用類比陣列3701中的金屬層M1至M4，M2至M4亦連接至週邊電路(例如，行解碼器)，而數位陣列位元線BLIFR0僅使用金屬層M1。類比陣列3701及數位陣列3704係同一物理陣列3907的一部分。物理陣列3907包括基板3908、擴散層3909及擴散層3910。基板3908由類比陣列3701及數位陣列3704共享。擴散層3909係類比陣列3701的一部分，但不是數位陣列3704的一部分，而擴散層3910係數位陣列3704的一部分，但不是類比陣列3701的一部分。物理陣列3907藉由使用用於類比陣列3701及數位陣列3704之斷開的金屬內連線及個別擴散層來分割。例如，可以看到，用於類比陣列3701的位元線BL0及用於數位陣列3704的位元線BLIFR0彼此斷開，並且類比陣列3701的擴散層3909與數位陣列3704的擴散層3910分開。類比陣列3701及數位陣列3704使用相同的物理陣列3907可減少因使用個

別的物理陣列時會發生之製程一致性及面積負擔所造成的任何物理效果。

【0154】 在圖39C中，VMM系統3903包括類比陣列3601及數位陣列3604。示例性位元線BL0耦接至類比陣列3601，而示例性位元線BLIFR0耦接至數位陣列3604。在此，位元線BL0及位元線BLIFR0係同一條位元線且都使用金屬層M1至M4。金屬層M2至M4亦連接至週邊電路(例如，行解碼器)。類比陣列3601及數位陣列3604係同一物理陣列3911的一部分。物理陣列3911包括基板3912及擴散層3913，它們均由類比陣列3601及數位陣列3604共享。與圖39A及39B中的物理陣列3904及3907不同，物理陣列3911沒有被分割。例如，可以看到，位元線BL0及BLIFR0係相同的且連接至類比陣列3601及數位陣列3604。類比陣列3601及數位陣列3604使用相同的物理陣列3911可減少因使用個別的物理陣列時會發生之製程一致性及面積負擔所造成的任何物理效果。

【0155】 圖40描繪示例VMM系統4000，其包括類比陣列4001、數位陣列4002(其可選地包括使用者資料陣列4008及系統資料陣列4009)、數位陣列4003(其可選地包括使用者資料陣列4006及系統資料陣列4007)、行解碼器4004及行解碼器4005。行解碼器4004耦接至與類比陣列4001的行耦接之位元線BL0、...、BLn。行解碼器4005耦接至與數位陣列4002及數位陣列4003的行耦接之位元線BLD0、...、BLDn。在此，數位陣列4003及數位陣列4002中的每一個陣列都可以用於儲存任何類型的數位資料，其包括用於非揮發性儲存操作之數位使用者資料。類比陣列4001及數位陣列4002、4003具有個別的行解碼器，可減少洩漏並提

高效能(如更快的速度)。可選地，類比陣列4001、數位陣列4003及數位陣列4002中的兩個或更多個被製造在同一半導體晶粒上。

【0156】 圖41描繪數位陣列4100，其可以是先前討論的數位陣列3604、3704、4002或4003中的任何一個。在此實例中，數位陣列4100包括使用者資料陣列4101及系統資料陣列4102。使用者資料陣列4101及系統資料陣列4102可以彼此獨立地進行存取，使得：(1)使用者資料陣列4101以平均頻率 f_1 來進行存取；而系統資料陣列4102以平均頻率 f_2 來進行存取，其中 f_1 與 f_2 可以是不同的， f_2 可以小於 f_1 ；(2)使用者資料陣列4101的存取時間為 t_1 ，而系統資料陣列4102的存取時間為 t_2 ，其中 t_1 與 t_2 可以是不同的， t_2 可以大於 t_1 ，這意味著讀取或寫入操作對於系統資料陣列4102來說比使用者資料陣列4101慢；(3)使用者資料陣列4101的耐久性為 e_1 ，而系統資料陣列4102的耐久性為 e_2 ，其中 e_1 與 e_2 可以是不同的， e_2 可以小於 e_1 。與使用者資料陣列4101相比，這可以允許系統資料陣列4102有更大的製造公差及更低的功率消耗。

【0157】 圖42A描繪差分電流至電壓轉換器4201，其可以在VMM系統3600、3700、4000中使用第一組電壓分別從類比陣列3601、3701及4001讀取類比資料及使用第二組電壓分別從數位陣列3604、3704以及4002及4003讀取數位資料。

【0158】 差分電流至電壓轉換器4201包括運算放大器4203；可變積分電阻器4204及4205；共模電路4206(用於運算放大器4203的差分放大器實現)。差分電流至電壓轉換器4201將兩個電流輸入 $IBL+$ 及 $IBL-$ 轉換成差分輸出電壓 $VO+$ 及 $VO-$ ，其中輸出電壓與可變電阻器4204及

4205的電阻成比例關係。輸入電流 $IBL+$ 及 $IBL-$ 可選地是代表正權重及負權重的電流。例如， $IBL+$ 可以是來自單個單元的電流 $Iw+$ 或者是來自耦合至位元線之複數個 $w+$ 單元的電流之和的位元線電流，而 $IBL-$ 可以是來自單個單元的電流 $Iw-$ 或者是來自耦接至位元線之複數個 $w-$ 單元的電流之和的位元線電流。這樣的正權重及負權重可以用在神經網路中來呈現一個權重($W=W+-W-$)。在另一個實例中，兩個輸入電流 $IBL+$ 及 $IBL-$ 可以表示單元電流(在這種情況下，它用於驗證權重調整中的單元電流目標，這意味著將單元程式化或抹除至目標電流)或來自陣列的位元線電流及參考電流。

【0159】 可選地，可以針對數位讀取將運算放大器4203及/或共模電路4206的偏壓電流設定為比針對類比讀取更高的電流位準。

【0160】 可選地，可以針對數位讀取與類比讀取將電阻器4204及4205設定為不同的數值。可選地，當轉換器4201用於權重調整操作中的單元驗證(例如，將記憶體單元程式化或抹除至目標電流)時，可以將電阻器4204及4205設定為不同的數值。圖42B描繪差分逐漸近似暫存器(SAR)類比至數位轉換器(ADC)4202，其可以在VMM系統3600、3700、4000中使用第一組電壓分別從類比陣列3601、3701及4001讀取類比資料以及使用第二組電壓分別從數位陣列3604、3704以及4002及4003讀取數位資料。

【0161】 差分逐漸近似暫存器類比至數位轉換器4202在所有可能的量化位準中使用二元搜索來識別適當的數位輸出，以將類比輸入或差分類比輸入轉換成數位輸出。

【0162】 差分逐漸近似暫存器類比至數位轉換器4202包括二進制電容數位至類比轉換器(CDAC)4207、二進制CDAC 4208(與CDAC 4207互補)、比較器4209及SAR邏輯及暫存器4210。

【0163】 差分逐漸近似暫存器類比至數位轉換器(SAR ADC)4202接收差分電壓輸入 V_{inp} 及 V_{inn} ，它們例如由差分電流至電壓轉換器4201提供。SAR邏輯及暫存器4210循環通過所有可能的數位位元組合，進而控制CDAC 4207及4208中的開關，以將電壓源耦接至電容器。當比較器4209的輸出翻轉時，將SAR邏輯及暫存器4210中的數位位元組合輸出作為數位輸出。可選地，SAR邏輯及暫存器4210在數位輸出中產生附加的1位元數位輸出DMAJ，其中如果數位值中的大多數位元為「1」，則附加的1位元數位輸出DMAJ為「1」；如果對應的數位值中之大多數位元不是「1」，則附加的1位元數位輸出DMAJ為「0」。

【0164】 可選地，電阻器4204及4205可以針對數位讀取操作設定為與類比讀取操作不同的電阻值。

【0165】 應當注意，如本文所使用，術語「在...上方」及「在...上」均包含性地包括「直接在...上」(沒有中間材料、元件或空間設置在其間)及「間接在...上」(中間材料、元件或空間設置在其間)。同樣地，術語「相鄰」包括「直接相鄰」(沒有中間材料、元件或空間設置在其間)及「間接相鄰」(中間材料、元件或空間設置在其間)，「安裝至」包括「直接安裝至」(沒有中間材料、元件或空間設置在其間)及「間接安裝至」(中間材料、元件或空間設置在其間)，以及「電耦接至」包括「直接電耦接至」(沒有中間材料或元件在其間將元件電連接在一起)及「間接電耦接至」(中間材料或元件在其間將元件電連接在一起)。

至」(中間材料或元件在其間將元件電連接在一起)。例如，「在基板上方」形成元件可以包括在基板上直接形成元件而在其間沒有中間材料/元件，以及在基板上間接形成元件而在其間具有一個或多個中間材料/元件。

【符號說明】

【0166】

12:半導體基板

14:源極區域

16:汲極區域

18:通道區域

20:浮動閘極

22:字元線端子(選擇閘極)

24:位元線

28:控制閘極

30:抹除閘極

31:數位至類比轉換器

32:向量矩陣乘法(VMM)陣列

32a:VMM陣列

32b:VMM陣列

32c:VMM陣列

32d:VMM陣列

32e:VMM陣列

33:非揮發性記憶體單元陣列

- 34:抹除閘極及字元線閘極解碼器
- 35:控制閘極解碼器
- 36:位元線解碼器
- 37:源極線解碼器
- 38:差分加法器
- 39:激勵函數區塊
- 210:記憶體單元
- 310:4閘極記憶體單元
- 410:3閘極記憶體單元
- 510:堆疊式閘極記憶體單元
- 900:神經元VMM陣列
- 901:非揮發性記憶體單元的記憶體陣列
- 902:非揮發性參考記憶體單元的參考陣列
- 903:控制閘極線
- 904:抹除閘極線
- 1000:神經元VMM陣列
- 1001:第一非揮發性參考記憶體單元的參考陣列
- 1002:第二非揮發性參考記憶體單元的參考陣列
- 1003:非揮發性記憶體單元的記憶體陣列
- 1014:多工器
- 1100:神經元VMM陣列
- 1101:第一非揮發性參考記憶體單元的參考陣列

- 1102:第二非揮發性參考記憶體單元的參考陣列
- 1103:非揮發性記憶體單元的記憶體陣列
- 1200:神經元VMM陣列
- 1201:第一非揮發性參考記憶體單元的參考陣列
- 1202:第二非揮發性參考記憶體單元的參考陣列
- 1203:非揮發性記憶體單元的記憶體陣列
- 1204:疊接電晶體
- 1205:多工器
- 1212:多工器
- 1300:神經元VMM陣列
- 1301:第一非揮發性參考記憶體單元的參考陣列
- 1302:第二非揮發性參考記憶體單元的參考陣列
- 1303:非揮發性記憶體單元的記憶體陣列
- 1314:多工器
- 1400:LSTM
- 1401:單元
- 1402:單元
- 1403:單元
- 1404:單元
- 1500:LSTM單元
- 1501:sigmoid函數裝置
- 1502:sigmoid函數裝置

1503:sigmoid函數裝置

1504:tanh裝置

1505:tanh裝置

1506:乘法裝置

1507:乘法裝置

1508:乘法裝置

1509:加法裝置

1600:LSTM單元

1601:VMM陣列

1602:激勵函數區塊

1700:LSTM單元

1701:VMM陣列

1702:激勵函數區塊

1703:乘法裝置

1704:暫存器

1705:暫存器

1706:暫存器

1707:暫存器

1708:加法裝置

1709:多工器

1710:多工器

1800:GRU

1801:單元

1802:單元

1803:單元

1804:單元

1900:GRU單元

1901:sigmoid函數裝置

1902:sigmoid函數裝置

1903:tanh裝置

1904:乘法裝置

1905:乘法裝置

1906:乘法裝置

1907:加法裝置

1908:互補裝置

2000:GRU單元

2001:VMM陣列

2002:激勵函數區塊

2100:GRU單元

2101:VMM陣列

2102:激勵函數區塊

2103:乘法裝置

2104:多工器

2105:加法裝置

2106:暫存器

2107:暫存器

2108:暫存器

2109:互補裝置

2200:神經元VMM陣列

2300:神經元VMM陣列

2400:神經元VMM陣列

2500:神經元VMM陣列

2600:神經元VMM陣列

2700:神經元VMM陣列

2701-1-2701-N:位元線控制閘

2800:神經元VMM陣列

2900:神經元VMM陣列

3000:神經元VMM陣列

3100:VMM系統

3101:求和電路

3102:求和電路

3210:VMM系統

3211:第一陣列

3212:第二陣列

3213:求和電路

3300:VMM系統

3301:陣列

3302:陣列

3303:求和電路

3304:求和電路

3305:求和電路

3306:求和電路

3307:求和電路

3308:求和電路

3400:習知技藝的VMM系統

3401:類比陣列

3402:列解碼器

3403:高電壓解碼器

3405:類比陣列區塊

3411:數位陣列

3412:列解碼器

3413:高電壓解碼器

3415:數位陣列區塊

3500:VMM系統

3501:VMM陣列

3502:列解碼器

3503:高電壓解碼器

3504:行解碼器

- 3505:位元線驅動器
- 3506:輸入電路
- 3507:輸出電路
- 3508:控制邏輯
- 3509:偏壓產生器
- 3510:高電壓產生區塊
- 3511:電荷泵
- 3512:電荷泵調節器
- 3513:高電壓類比精確位準產生器
- 3514:演算法控制器
- 3515:類比電路
- 3516:控制引擎
- 3517:測試控制邏輯
- 3600:VMM系統
- 3601:類比陣列
- 3602:列解碼器
- 3603:高電壓解碼器
- 3604:數位陣列
- 3605:列解碼器
- 3606:高電壓解碼器
- 3607:使用者資料陣列
- 3608:系統資料陣列

3700:VMM系統

3701:類比陣列

3702:列解碼器

3703:高電壓解碼器

3704:數位陣列

3705:列解碼器

3706:高電壓解碼器

3707:使用者資料陣列

3708:系統資料陣列

3801:行解碼器

3802:行解碼器

3901:VMM系統

3902:VMM系統

3903:VMM系統

3904:物理陣列

3905:基板

3906:擴散層

3907:物理陣列

3908:基板

3909:擴散層

3910:擴散層

3911:物理陣列

3912:基板

3913:擴散層

4000:VMM系統

4001:類比陣列

4002:數位陣列

4003:數位陣列

4004:行解碼器

4005:行解碼器

4006:使用者資料陣列

4007:系統資料陣列

4008:使用者資料陣列

4009:系統資料陣列

4100:數位陣列

4101:使用者資料陣列

4102:系統資料陣列

4201:差分電流至電壓轉換器

4202:差分逐漸近似暫存器(SAR)類比至數位轉換器(ADC)

4203:運算放大器

4204:可變積分電阻器

4205:可變積分電阻器

4206:共模電路

4207:二進制電容數位至類比轉換器(CDAC)

4208:二進制CDAC

4209:比較器

4210:SAR邏輯及暫存器

BL0:類比陣列位元線

BL0-BLN:位元線

BL0-BLn:位元線

BLD0-BLDn:位元線

BLIFR0:數位陣列位元線

BLR0:端子

BLR1:端子

BLR2:端子

BLR3:端子

c_0 :單元狀態向量

c_1 :單元狀態向量

c_2 :單元狀態向量

$c(t-1)$:單元狀態向量

$c(t)$:單元狀態向量

C1:層

C2:層

C3:層

CB1:突觸

CB2:突觸

- CB3:突觸
- CB4:突觸
- CG0:控制閘極線(電壓輸入)
- CG1:控制閘極線(電壓輸入)
- CG2:控制閘極線(電壓輸入)
- CG3:控制閘極線(電壓輸入)
- CG₀-CG_N:控制閘極線
- DMAJ:1位元數位輸出
- e1:耐久性
- e2:耐久性
- EG0:EG線
- EG1:EG線
- EGR0:EG線
- EGR1:EG線
- f₁:平均頻率
- f₂:平均頻率
- h₀:輸出向量
- h₁:輸出向量
- h₂:輸出向量
- h₃:輸出向量
- h(t-1):輸出向量
- h(t):輸出向量

IBL+: 電流輸入

IBL-: 電流輸入

INPUT₀-INPUT_M: 輸入

INPUT₀-INPUT_N: 輸入

M1-M4: 金屬層

OUTPUT₀-OUTPUT_N: 輸出

P1: 激勵函數

P2: 激勵函數

S0: 輸入層

S1: 層

S2: 層

S3: 輸出層

SL0: 源極線

SL1: 源極線

SL2: 源極線

SL3: 源極線

t₁: 存取時間

t₂: 存取時間

V_{inn}: 差分電壓輸入

V_{inp}: 差分電壓輸入

VO+: 差分輸出電壓

VO-: 差分輸出電壓

WL0:電壓輸入(字元線)

WL1:電壓輸入(字元線)

WL2:電壓輸入(字元線)

WL3:電壓輸入(字元線)

WL₀-WL_M:字元線

WLA0:字元線

WLA1:字元線

WLA2:字元線

WLA3:字元線

WLB0:字元線

WLB1:字元線

WLB2:字元線

WLB3:字元線

x₀:輸入向量

x₁:輸入向量

x₂:輸入向量

x₃:輸入向量

x(t):輸入向量

【發明申請專利範圍】

【請求項1】 一種人工神經網路系統，包括：

一類比陣列，其具有以列與行排列且配置用以儲存類比資料的非揮發性記憶體單元；

一數位陣列，其具有以列與行排列且配置用以儲存數位資料的非揮發性記憶體單元；以及

複數條位元線，其中該複數條位元線中的每條位元線連接至該類比陣列中的一行非揮發性記憶體單元及該數位陣列中的一行非揮發性記憶體單元。

【請求項2】 如請求項1之人工神經網路系統，包括：

一第一列解碼器，其耦接至該類比陣列；以及

一第一高電壓列解碼器，其耦接至該類比陣列。

【請求項3】 如請求項2之人工神經網路系統，包括：

一第二列解碼器，其與該第一列解碼器分開且耦接至該數位陣列；
以及

一第二高電壓列解碼器，其與該第一高電壓列解碼器分開且耦接至該數位陣列。

【請求項4】 如請求項3之人工神經網路系統，其中，在該類比陣列的讀取期間，該第一列解碼器及該第一高電壓列解碼器施加一第一組電壓至該類比陣列。

【請求項5】 如請求項4之人工神經網路系統，其中，在該數位陣列的讀取期間，該第二列解碼器及該第二高電壓列解碼器施加一第二組電

壓至該數位陣列，該第二組電壓不同於該第一組電壓。

【請求項6】 如請求項3之人工神經網路系統，包括在該類比陣列的讀取及該數位陣列的讀取期間使用的一讀取電路。

【請求項7】 如請求項6之人工神經網路系統，其中，在該數位陣列的讀取期間，如果該讀取電路產生的大多數輸出位元是「1」，則該讀取電路輸出「1」；如果該讀取電路產生的大多數輸出位元不是「1」，則該讀取電路輸出「0」。

【請求項8】 如請求項1之人工神經網路系統，其中，該數位陣列包括一使用者資料陣列及一系統資料陣列。

【請求項9】 如請求項8之人工神經網路系統，其中，該系統資料陣列的讀取、程式化或抹除中之一或多者的速度比對該使用者資料陣列所執行的慢。

【請求項10】 如請求項8之人工神經網路系統，其中，該系統資料陣列的讀取之速度比對該使用者資料陣列所執行的慢。

【請求項11】 如請求項8之人工神經網路系統，其中，該系統資料陣列的耐久性小於該使用者資料陣列的耐久性。

【請求項12】 一種人工神經網路系統，包括：

一類比陣列，其具有以列與行排列且配置用以儲存類比資料的非揮發性記憶體單元；

一數位陣列，其具有以列與行排列且配置用以儲存數位資料的非揮發性記憶體單元，該類比陣列與該數位陣列被製造在同一半導體晶粒上；

複數條第一位元線，其中該複數條第一位元線中的每條位元線耦接

至該類比陣列中之一行非揮發性記憶體單元；以及

複數條第二位元線，其中該複數條第二位元線中的每條位元線耦接至該數位陣列中之一行非揮發性記憶體單元，並且該複數條第二位元線與該複數條第一位元線斷開。

【請求項13】 如請求項12之人工神經網路系統，包括：

一第一列解碼器，其耦接至該類比陣列；以及

一第一高電壓列解碼器，其耦接至該類比陣列。

【請求項14】 如請求項13之人工神經網路系統，包括：

一第二列解碼器，其與該第一列解碼器分開且耦接至該數位陣列；

以及

一第二高電壓列解碼器，其與該第一高電壓列解碼器分開且耦接至該數位陣列。

【請求項15】 如請求項14之人工神經網路系統，其中，在該類比陣列的讀取期間，該第一列解碼器及該第一高電壓列解碼器施加一第一組電壓至該類比陣列。

【請求項16】 如請求項15之人工神經網路系統，其中，在該數位陣列的讀取期間，該第二列解碼器及該第二高電壓列解碼器施加一第二組電壓至該數位陣列，該第二組電壓不同於該第一組電壓。

【請求項17】 如請求項12之人工神經網路系統，其中，該複數條第一位元線及該複數條第二位元線共享一擴散層。

【請求項18】 如請求項17之人工神經網路系統，其中，該複數條第一位元線及該複數條第二位元線具有不同的金屬內連線。

【請求項19】 如請求項12之人工神經網路系統，其中，該複數條第一位元線及該複數條第二位元線具有不同的擴散層。

【請求項20】 如請求項18之人工神經網路系統，其中，該複數條第一位元線及該複數條第二位元線具有不同的金屬內連線。

【請求項21】 如請求項12之人工神經網路系統，其中，該數位陣列包括一系統資料陣列。

【請求項22】 如請求項12之人工神經網路系統，包括在該類比陣列的讀取及該數位陣列的讀取期間使用的一讀取電路。

【請求項23】 如請求項21之人工神經網路系統，其中，在該數位陣列的讀取期間，如果一讀取電路產生的大多數輸出位元是「1」，則該讀取電路輸出「1」；如果該讀取電路產生的大多數輸出位元不是「1」，則該讀取電路輸出「0」。

【請求項24】 如請求項12之人工神經網路系統，包括：
一第二數位陣列，其具有以列與行排列的非揮發性記憶體單元。

【請求項25】 如請求項24之人工神經網路系統，其中，該複數條第二位元線中的每條位元線耦接至該第二數位陣列中之一行非揮發性記憶體單元。

【請求項26】 如請求項12之人工神經網路系統，其中，該非揮發性記憶體單元數位陣列包括一使用者資料陣列及一系統資料陣列。

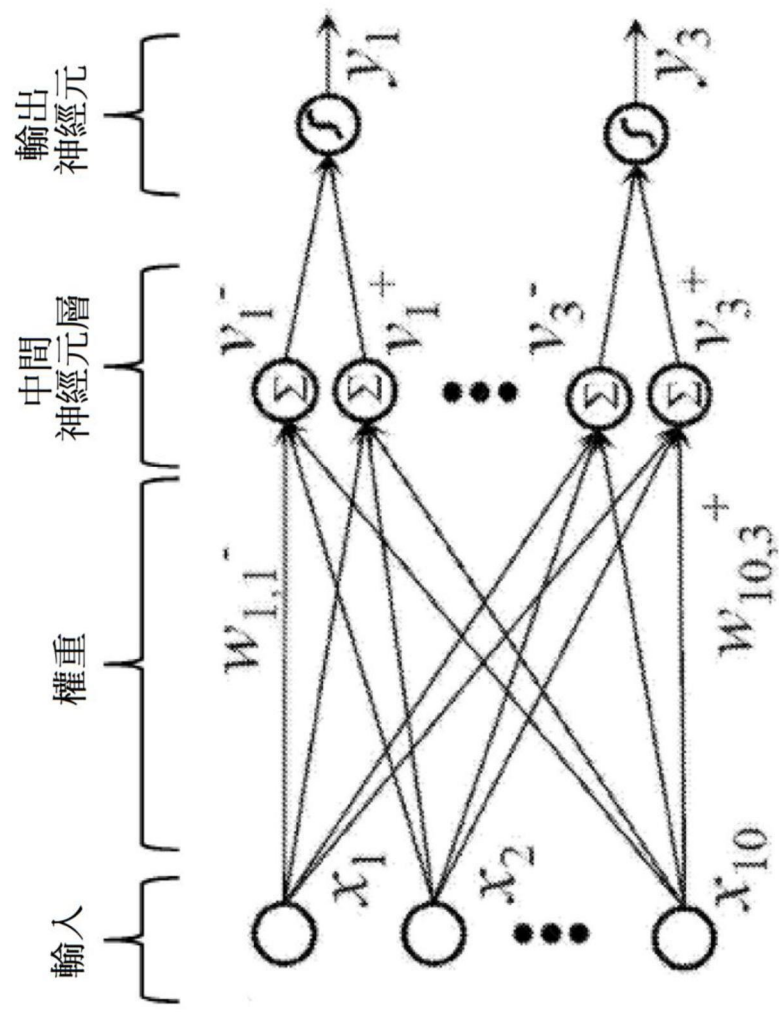
【請求項27】 如請求項26之人工神經網路系統，其中，該系統資料陣列的讀取、程式化或抹除操作中之一或多者的速度比對該使用者資料陣列所執行的慢。

2023.10.30 替換本

【請求項28】 如請求項26之人工神經網路系統，其中，該系統資料陣列的讀取操作之速度比對該使用者資料陣列所執行的慢。

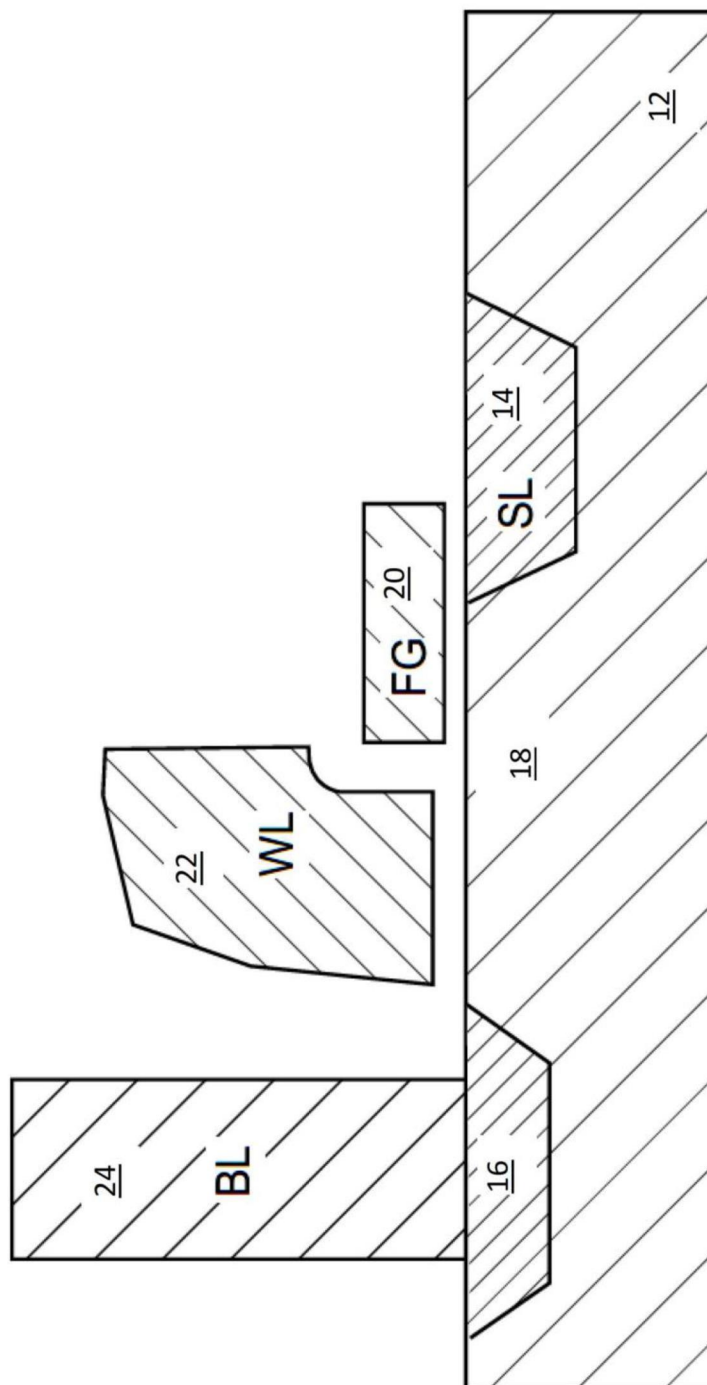
【請求項29】 如請求項26之人工神經網路系統，其中，該系統資料陣列的耐久性小於該使用者資料陣列的耐久性。

【發明圖式】



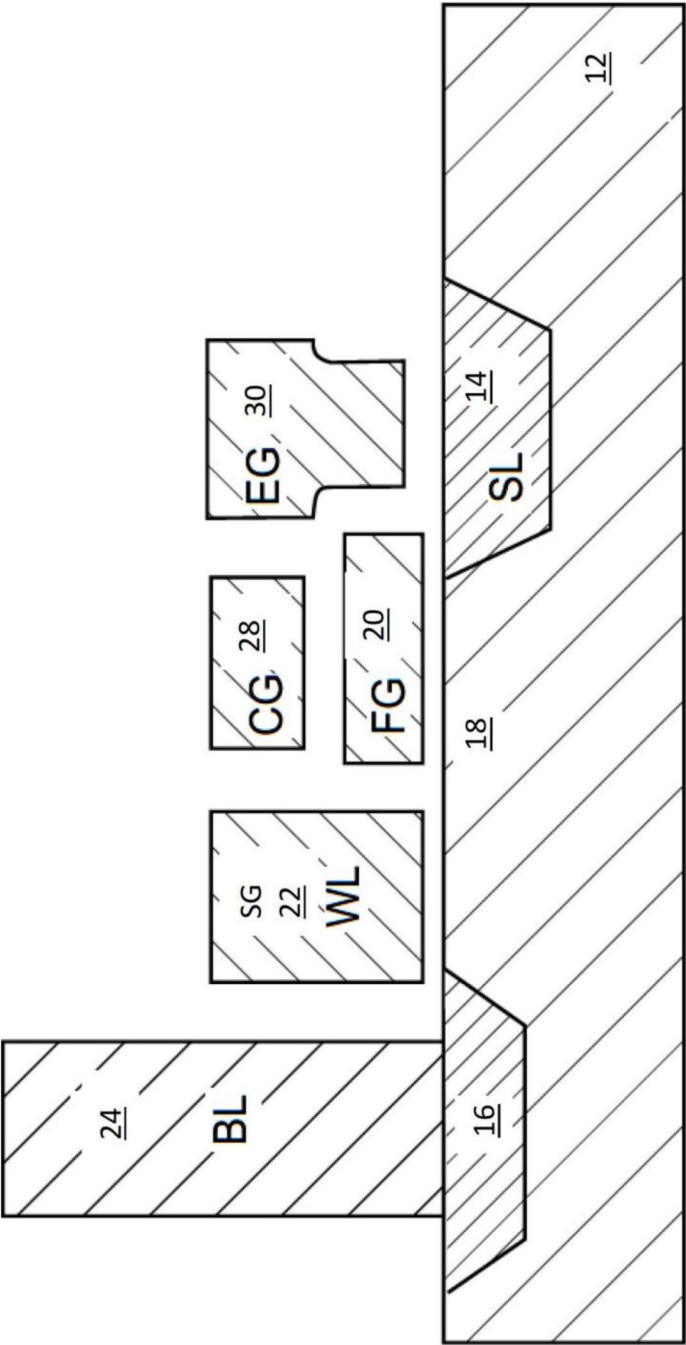
【圖1】

記憶體單元



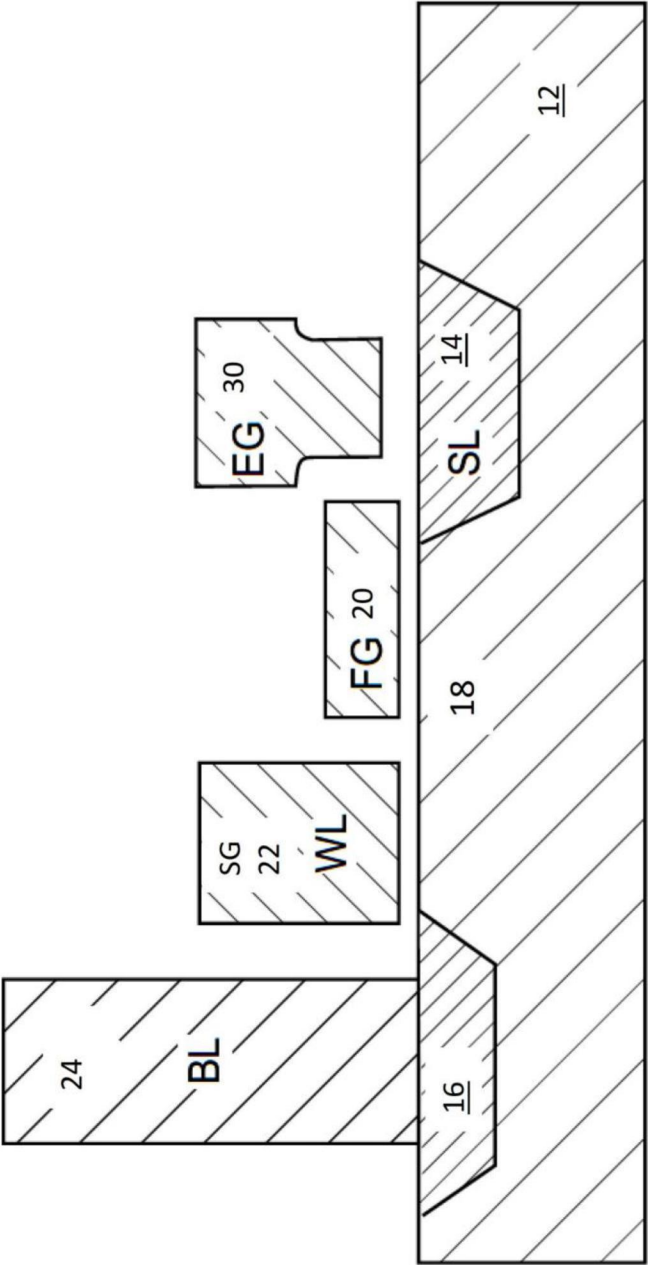
【圖2】

記憶體單元
310



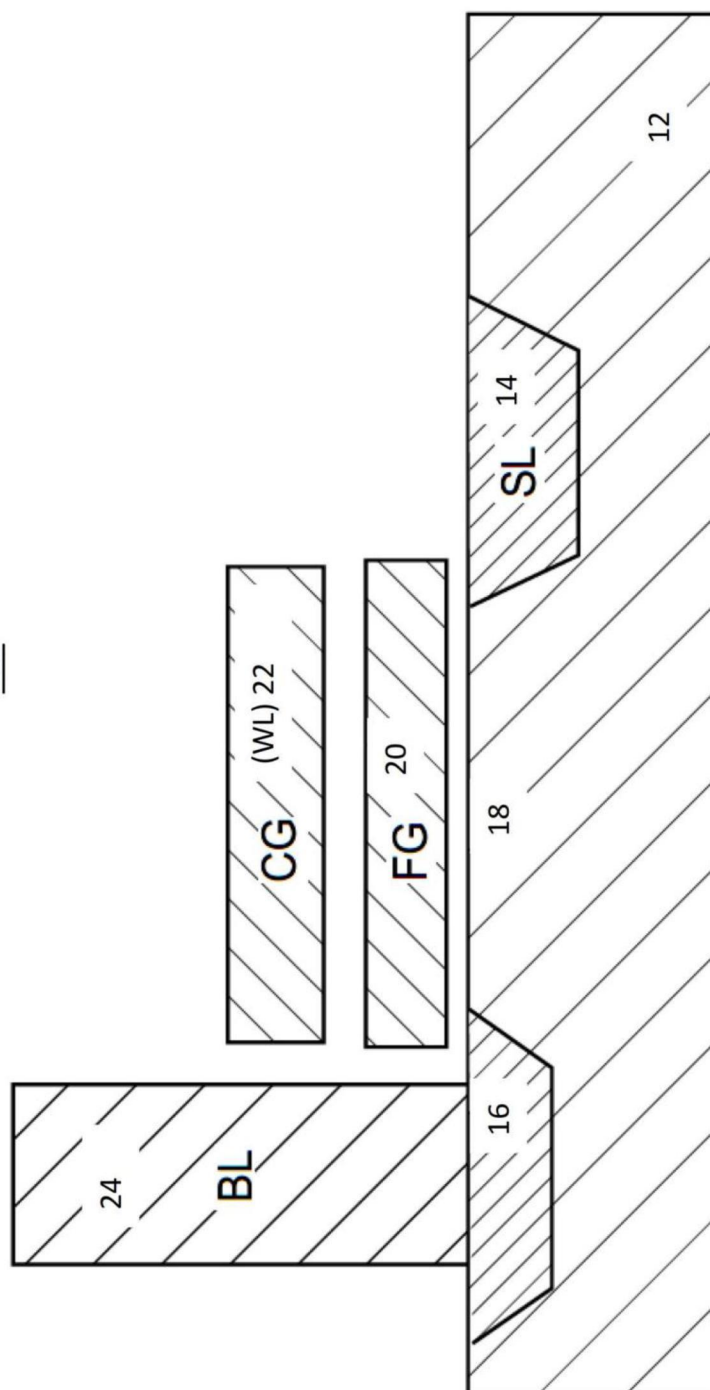
【圖3】

記憶體單元
410

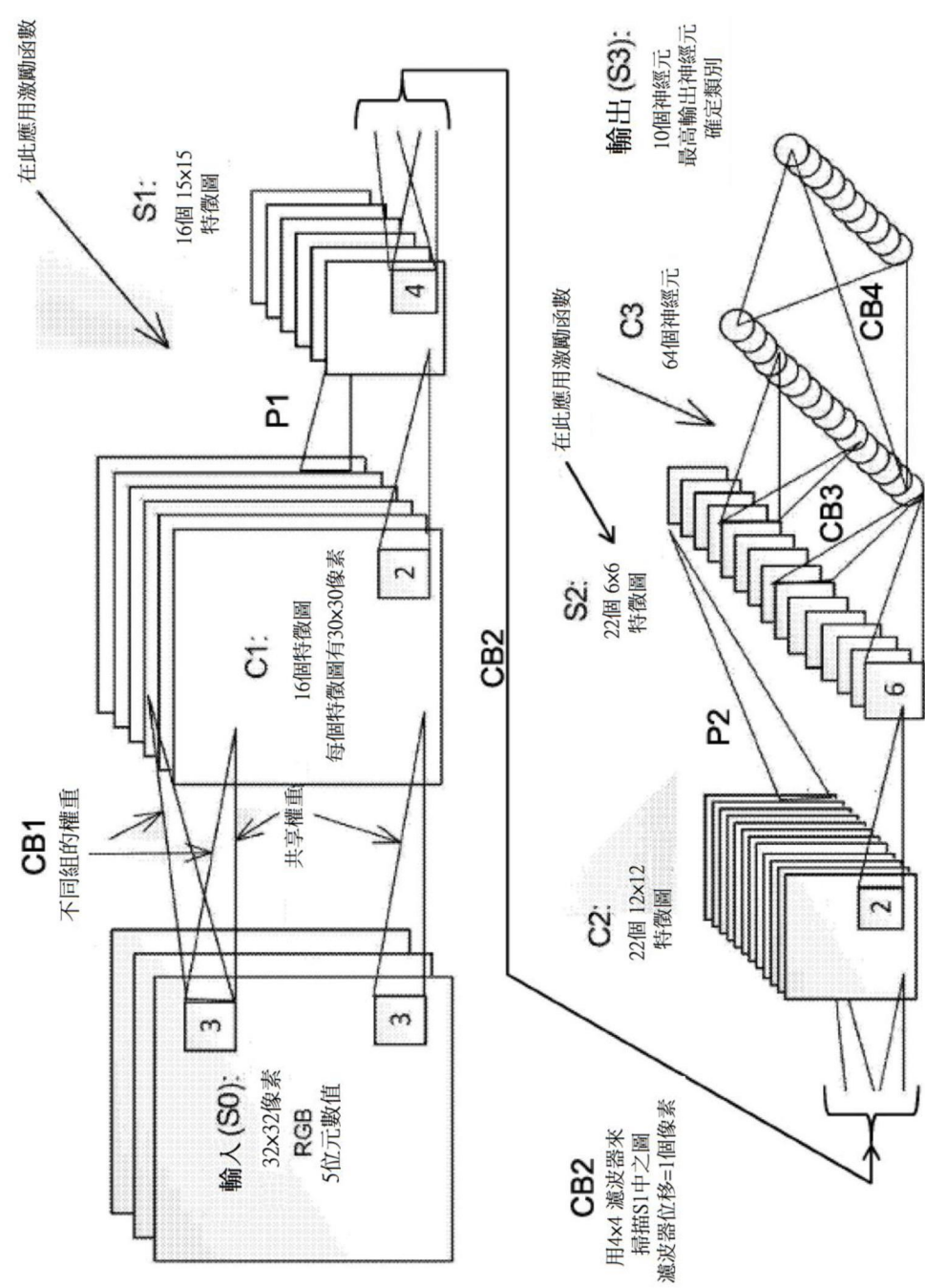


【圖4】

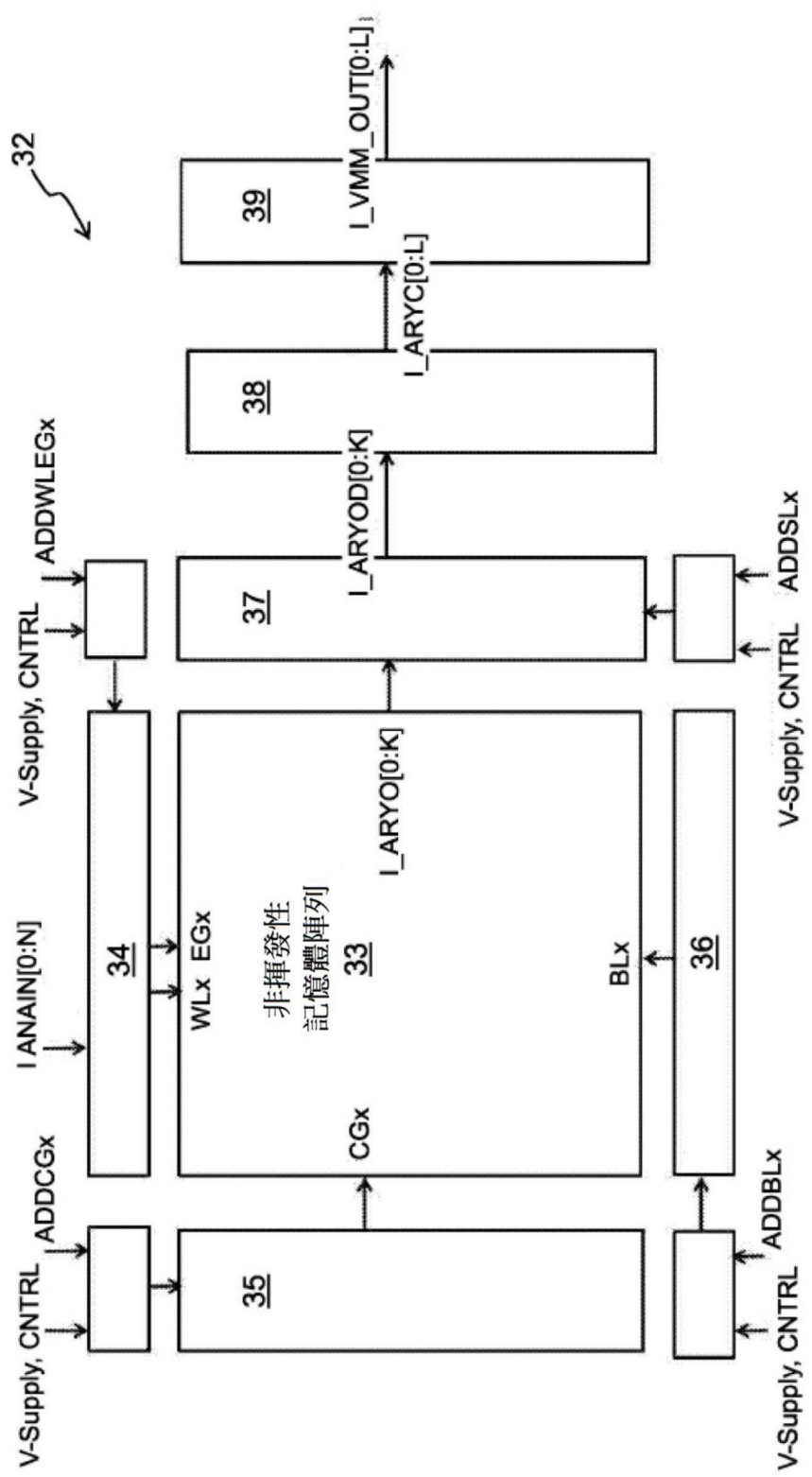
記憶體單元



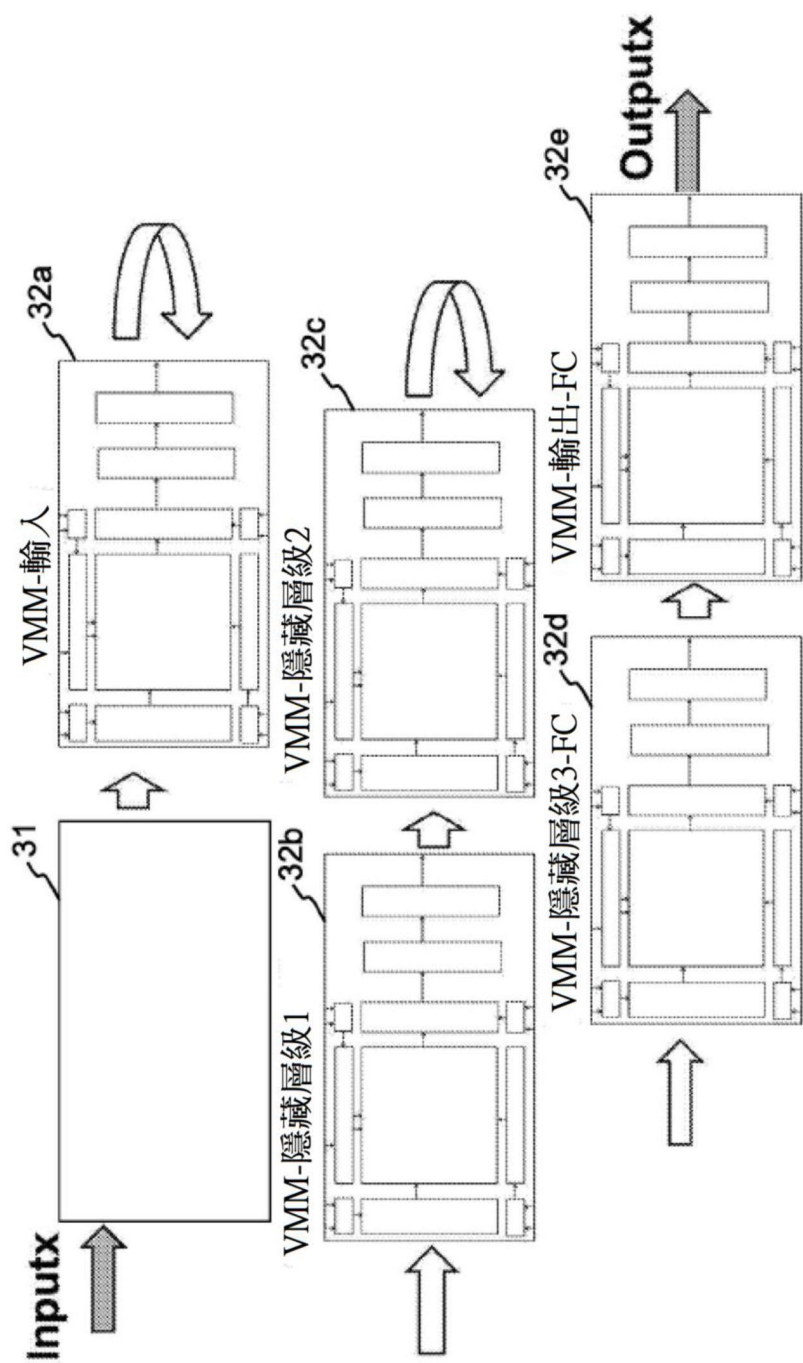
【5回】



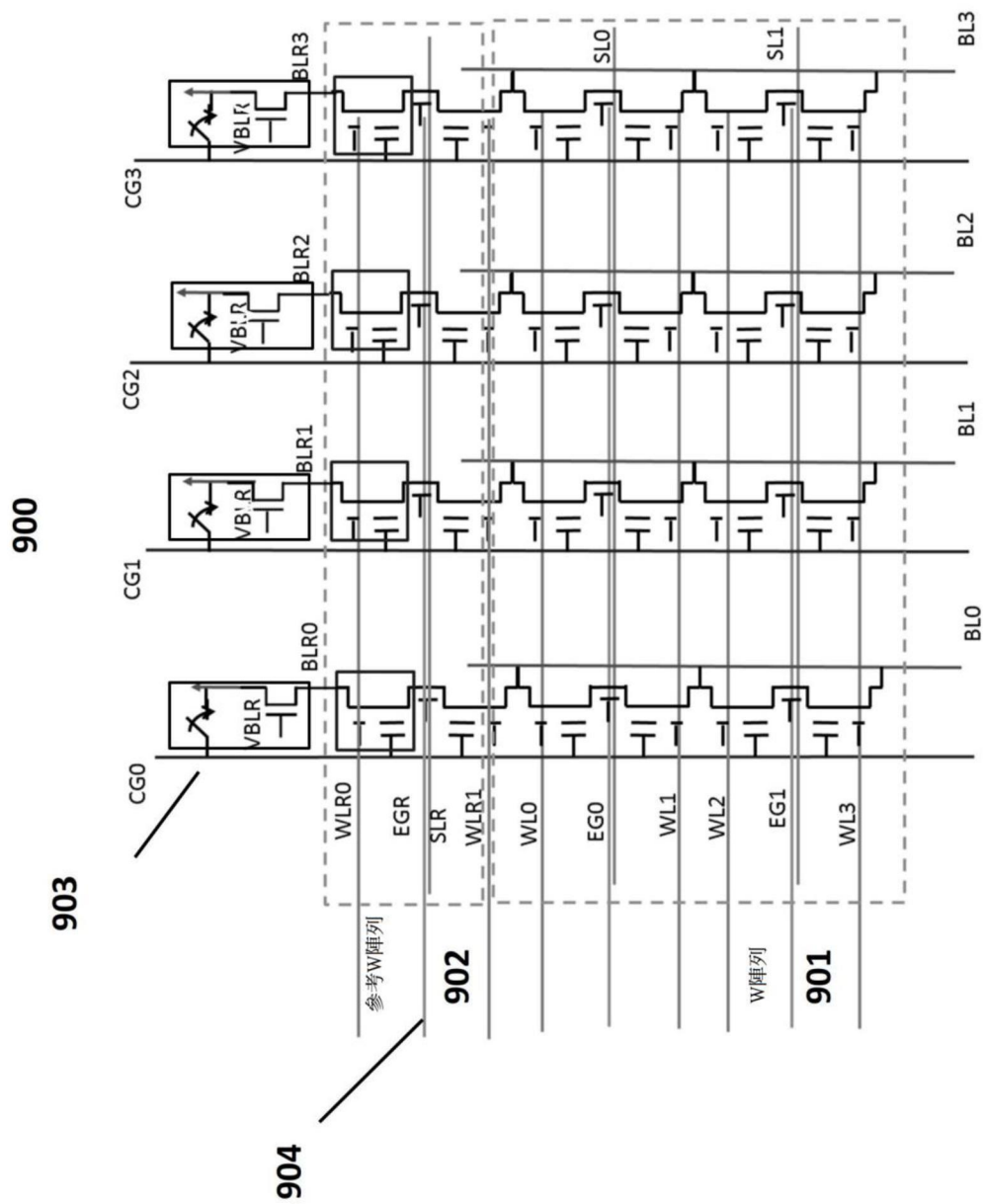
【圖6】



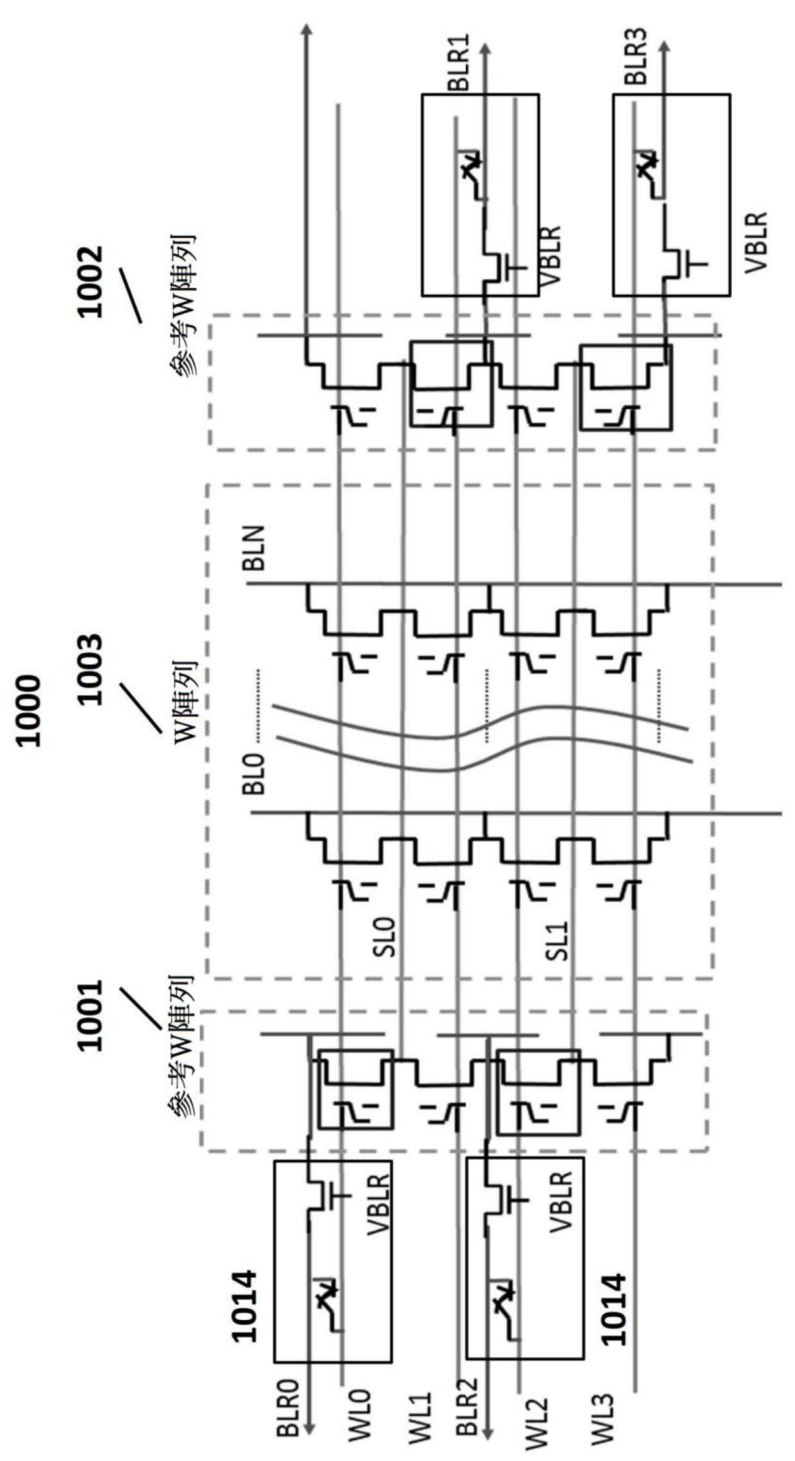
【圖7】



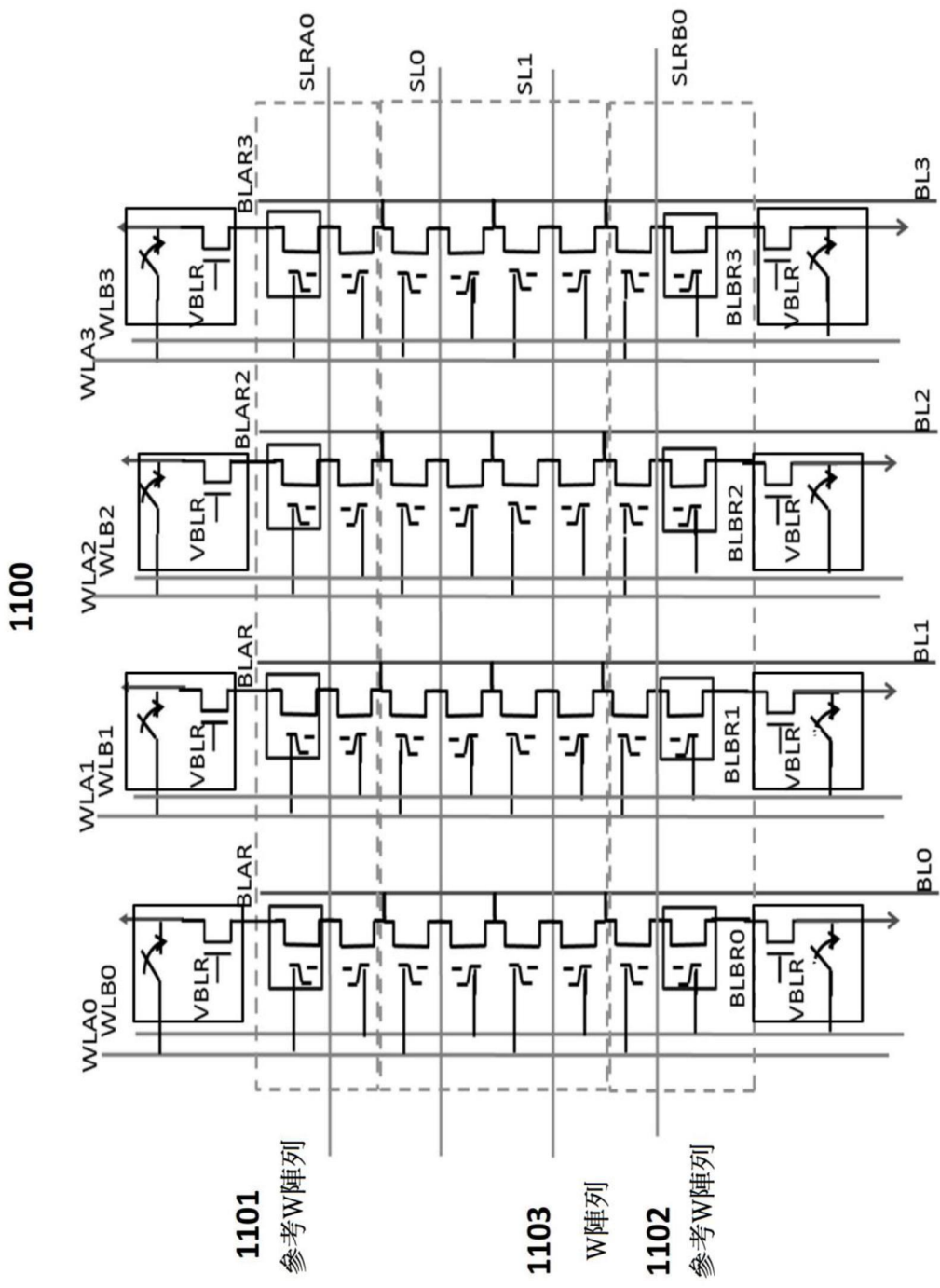
【圖8】



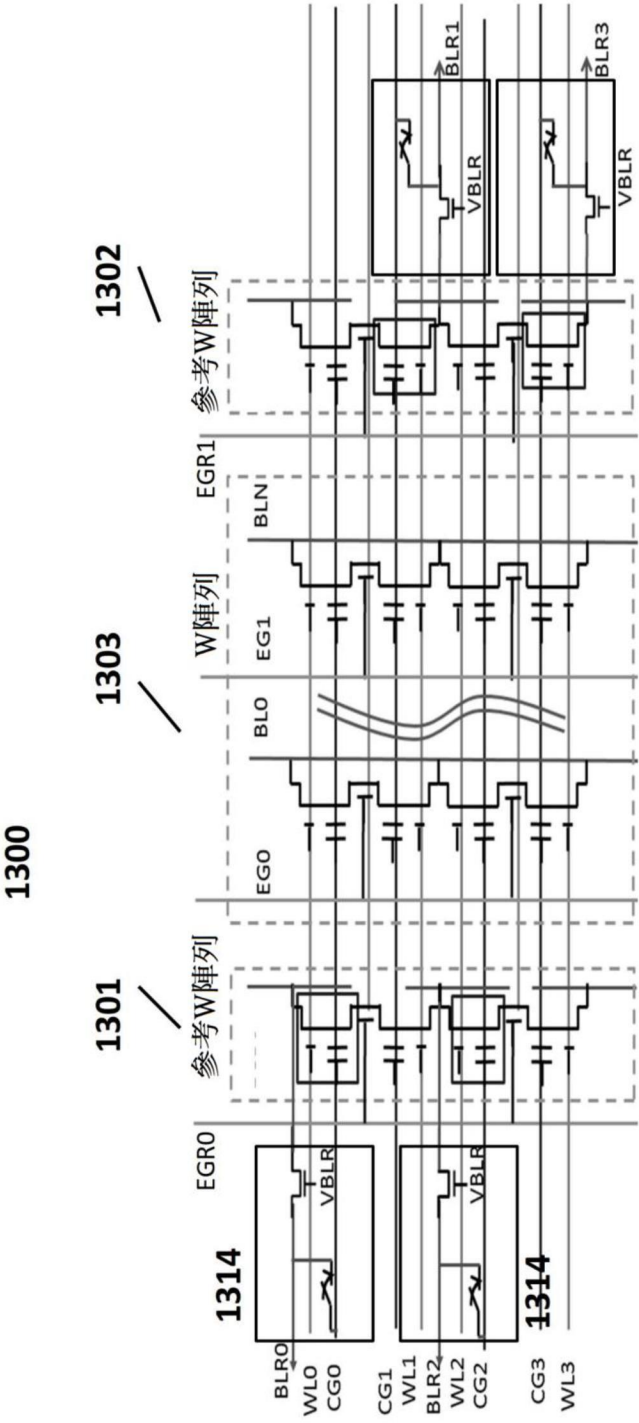
【圖9】



【圖10】

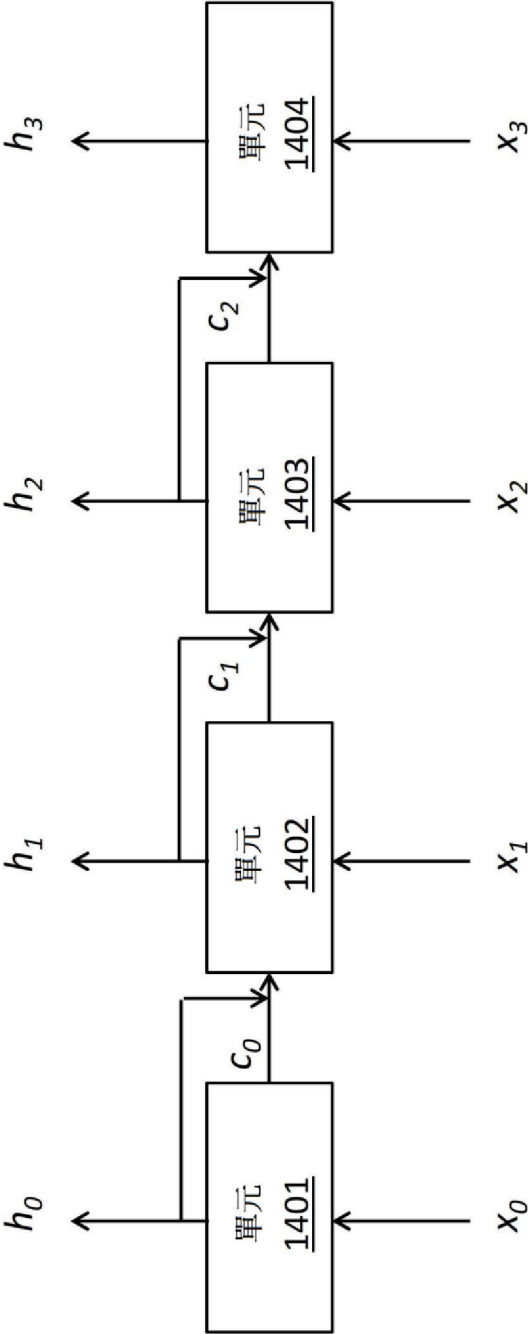


【圖11】



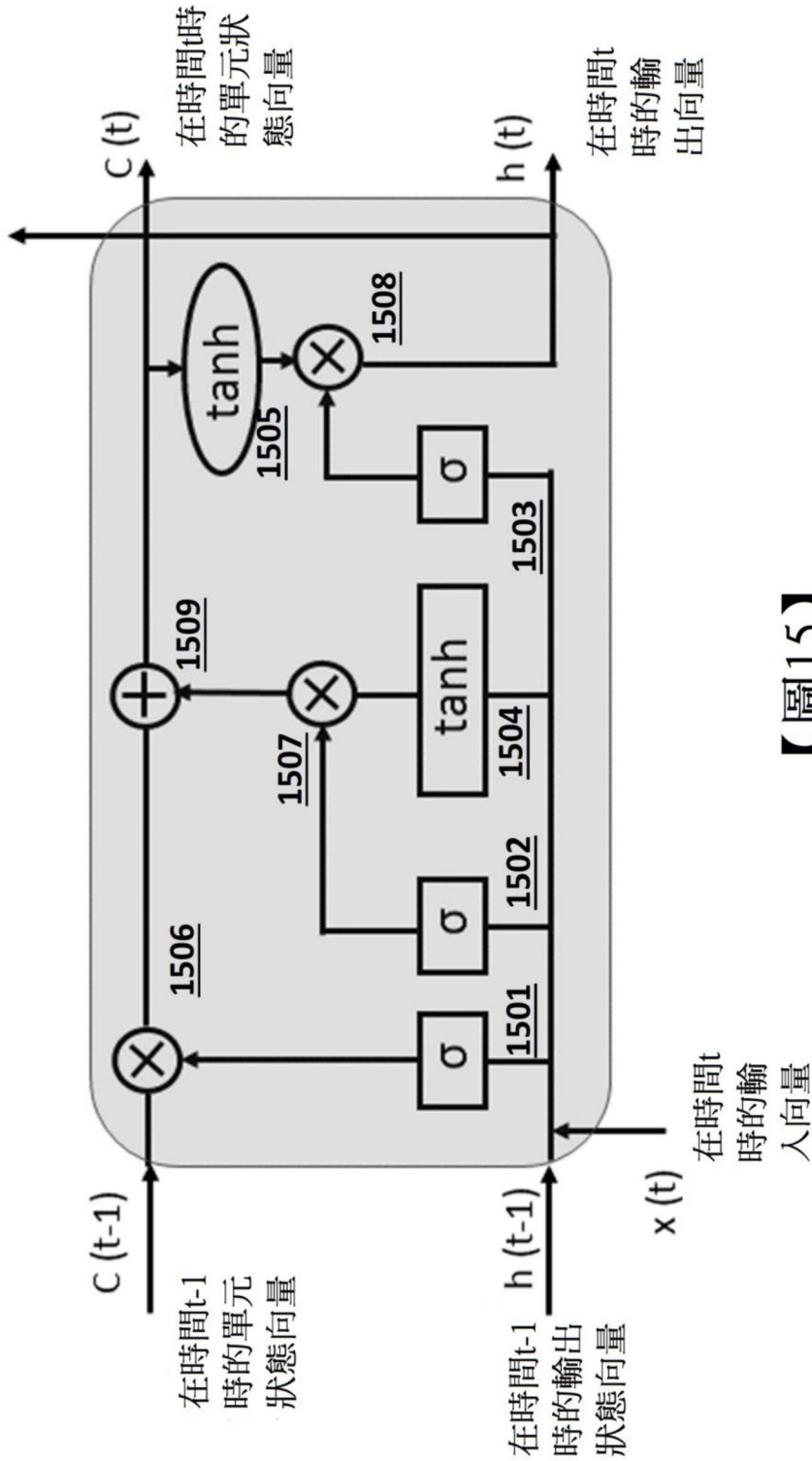
【圖13】

LSTM
1400

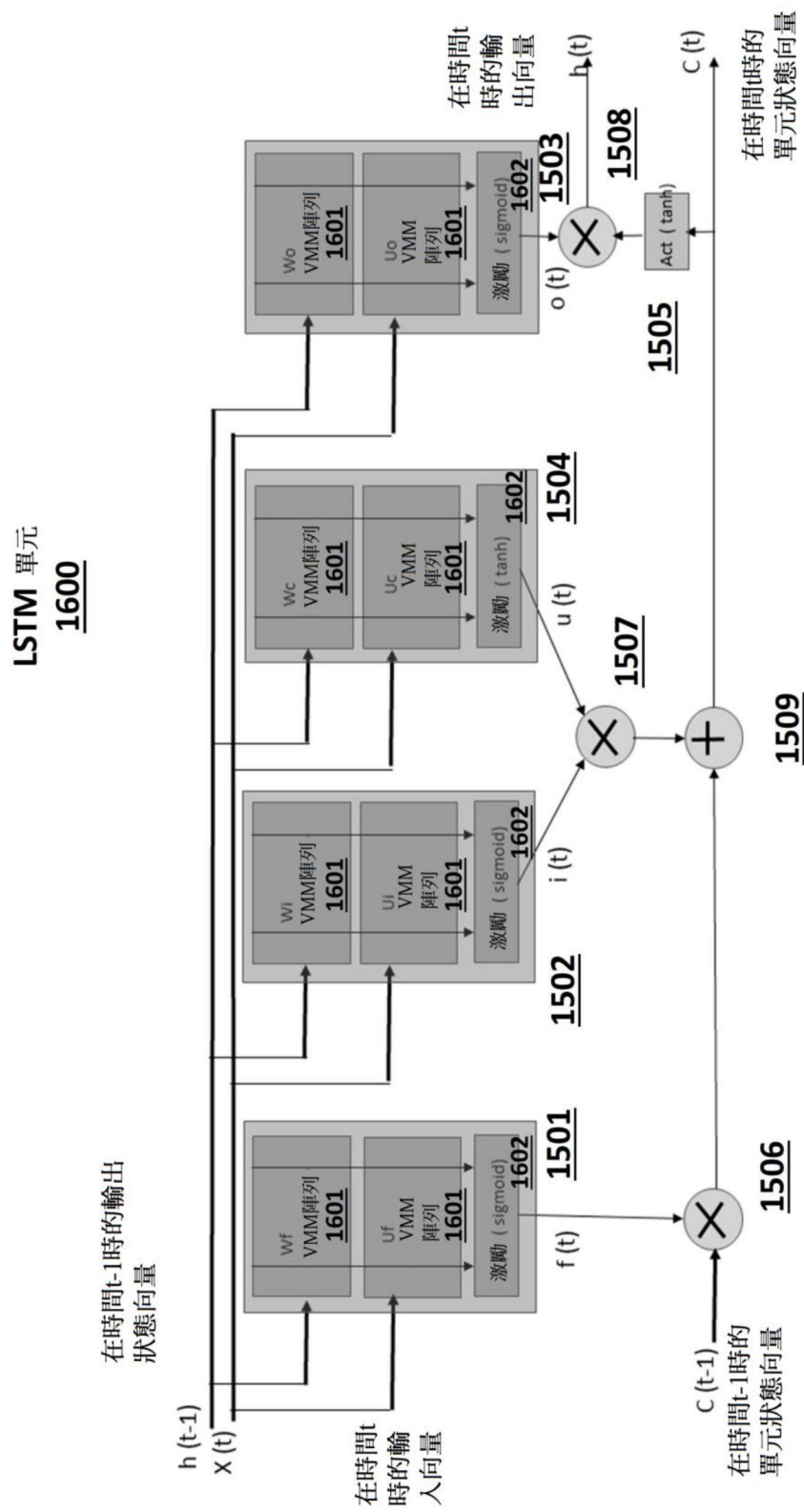


【圖14】

LSTM 單元
1500

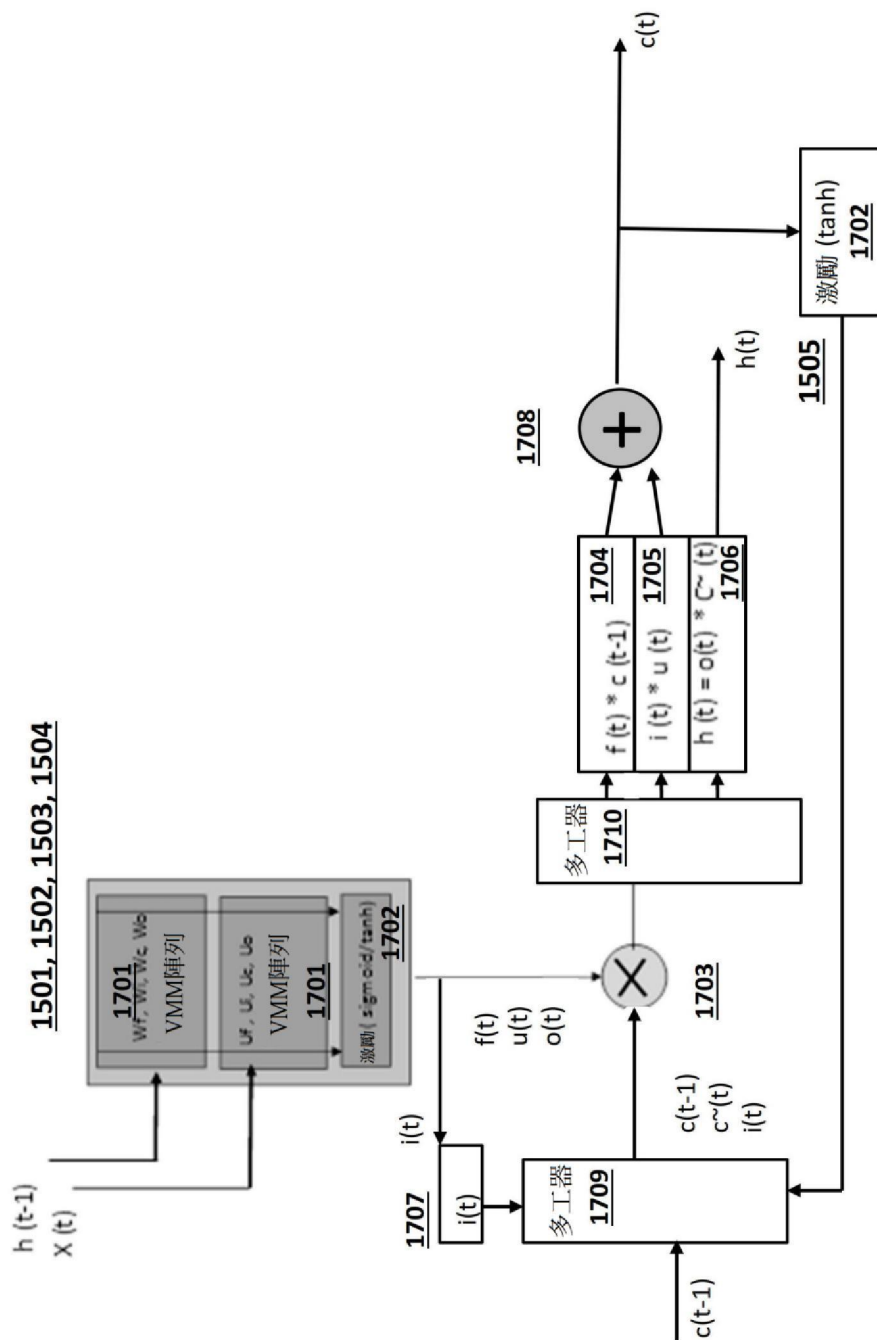


【圖15】



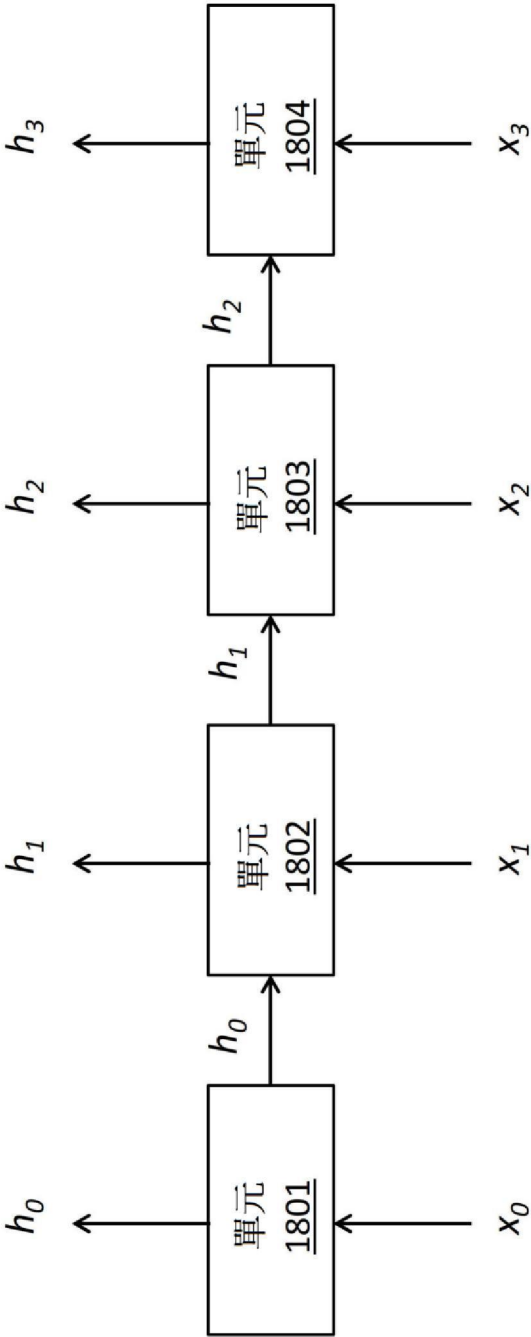
【圖16】

LSTM 單元
1700

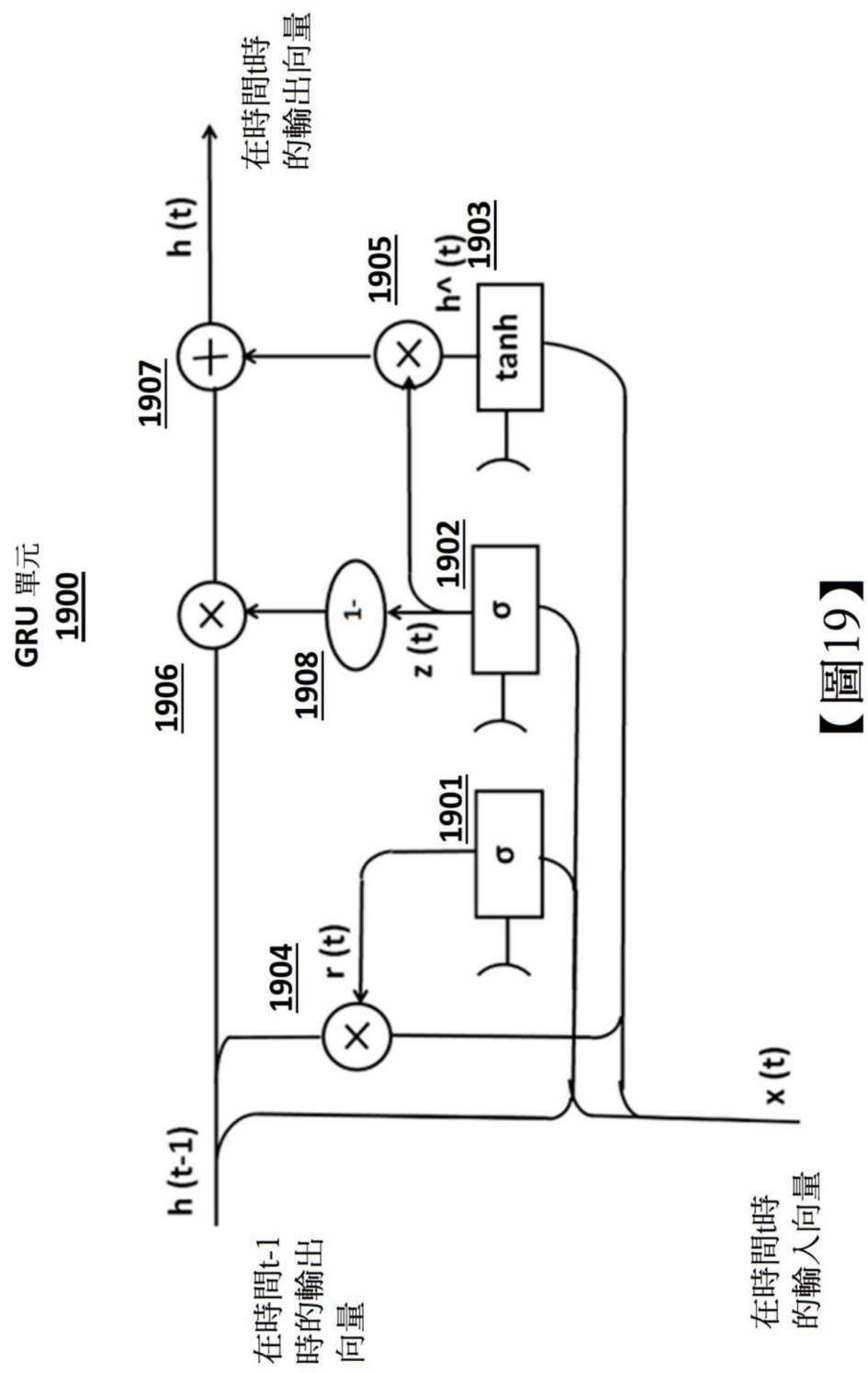


【圖17】

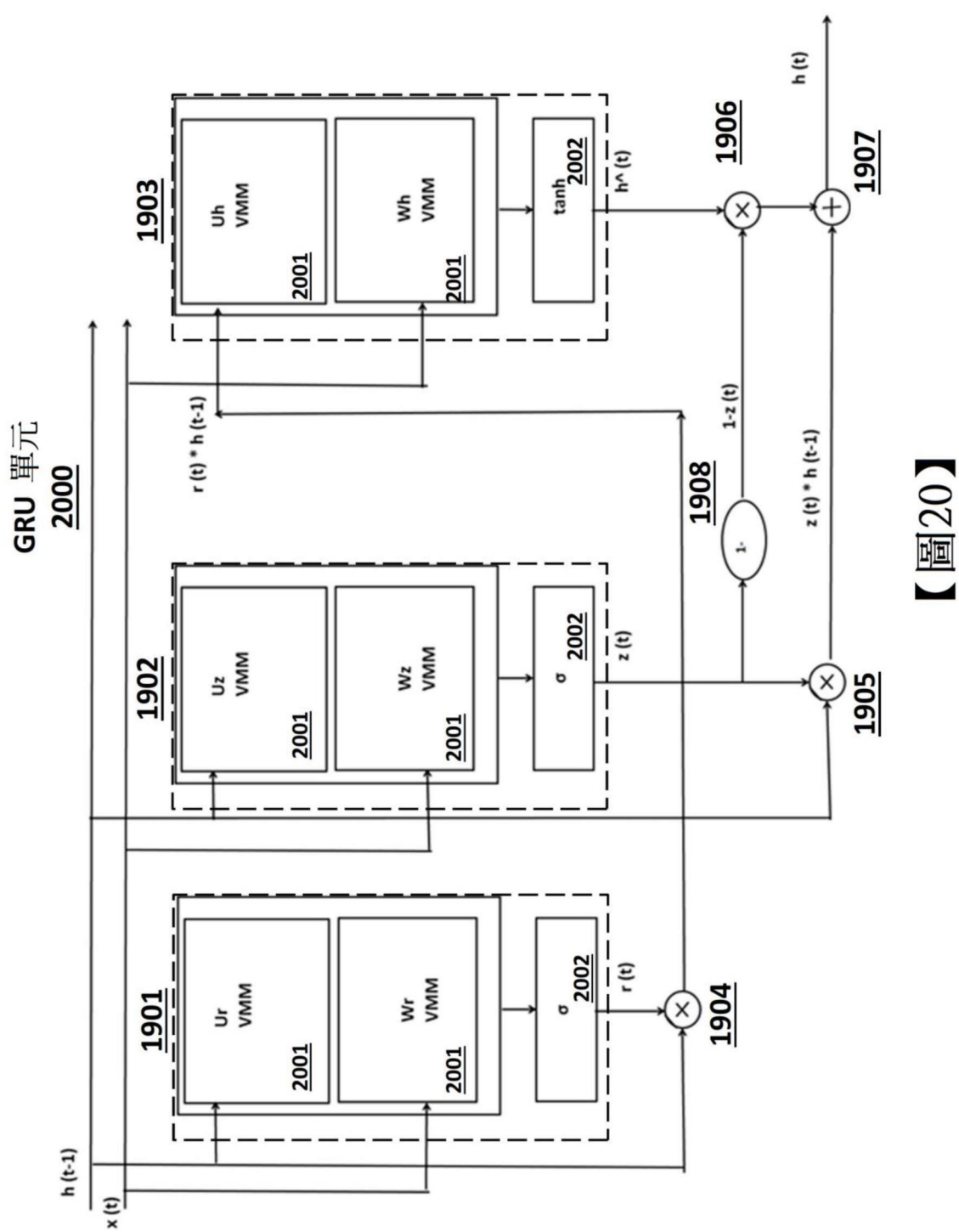
GRU
1800



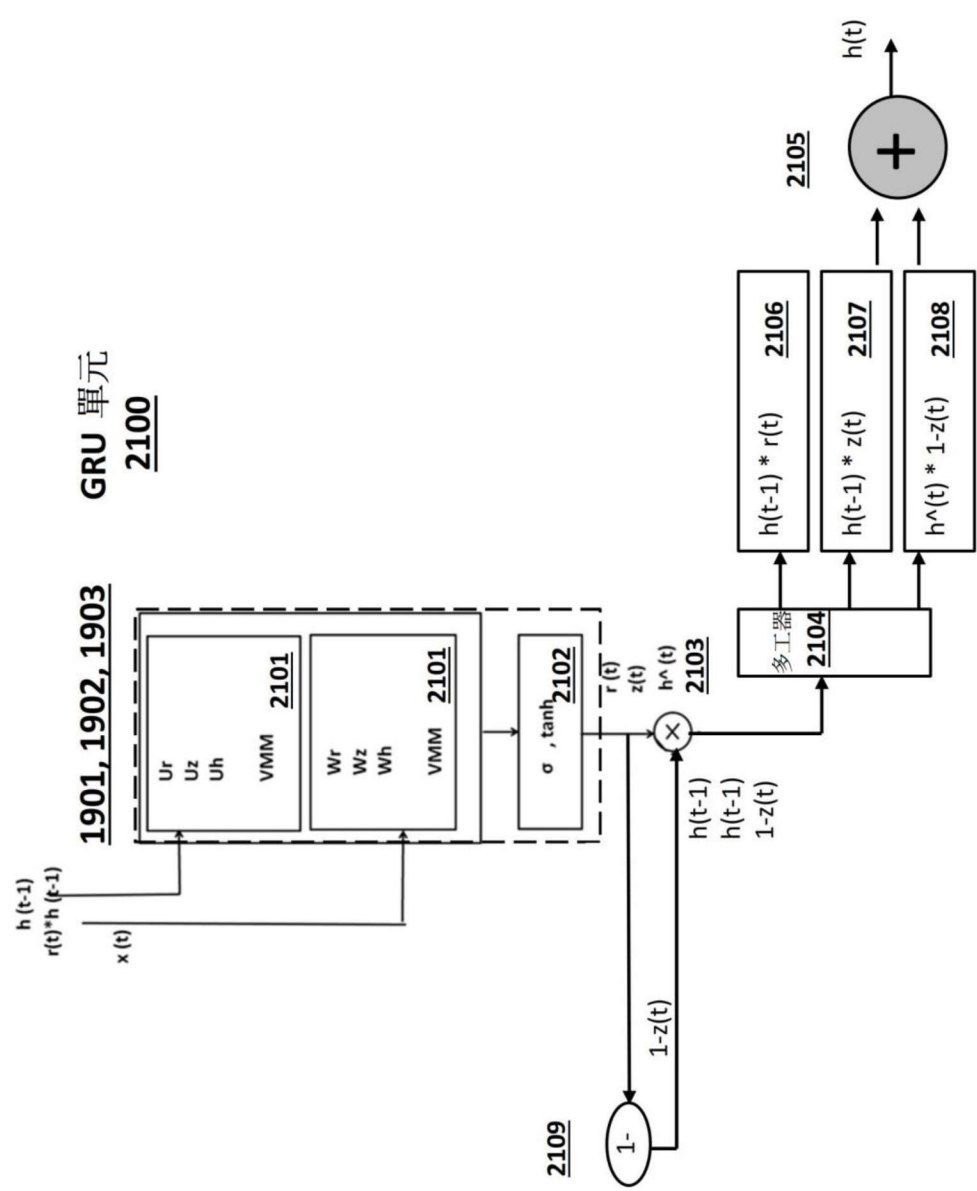
【圖18】



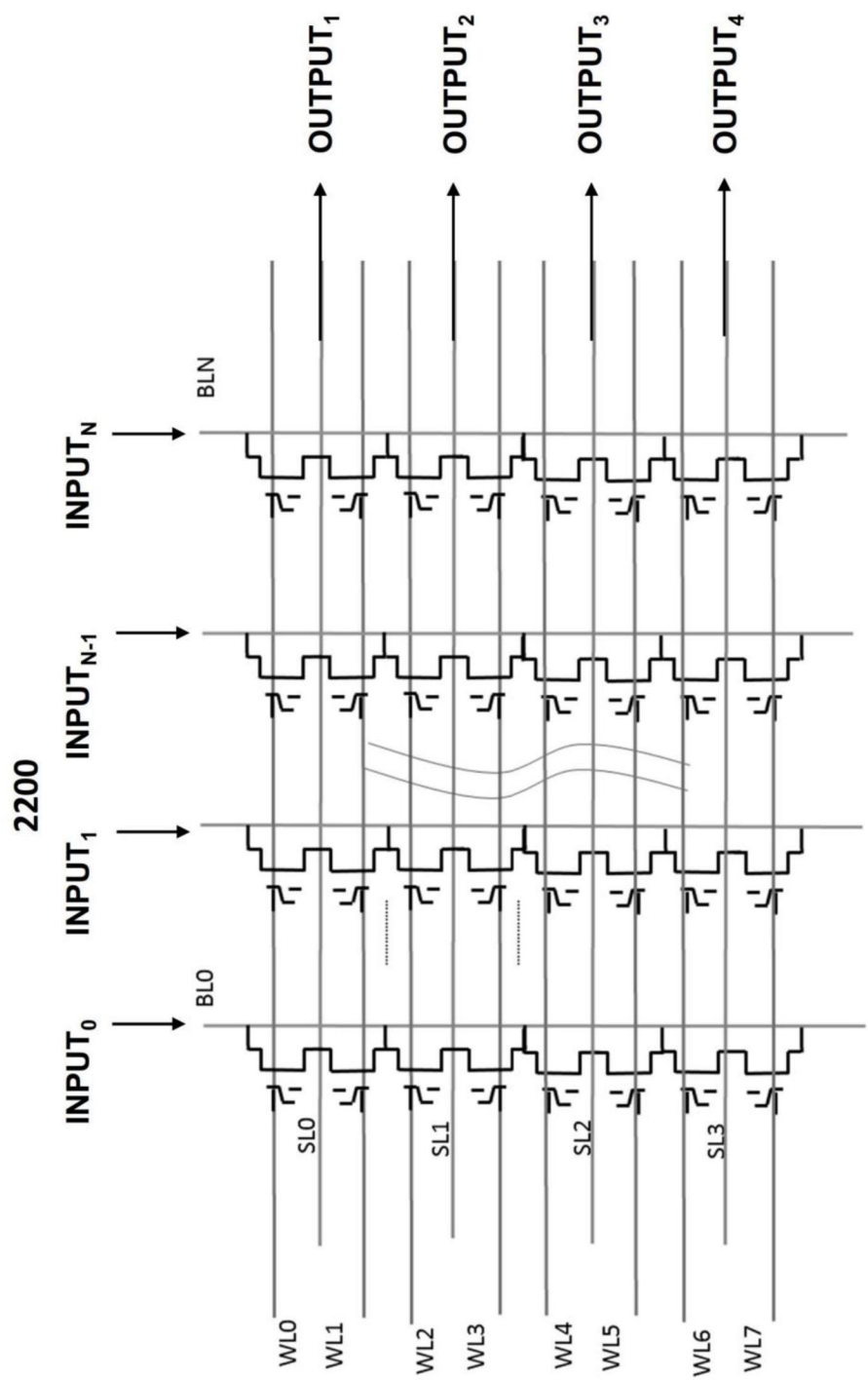
【圖19】



【圖20】

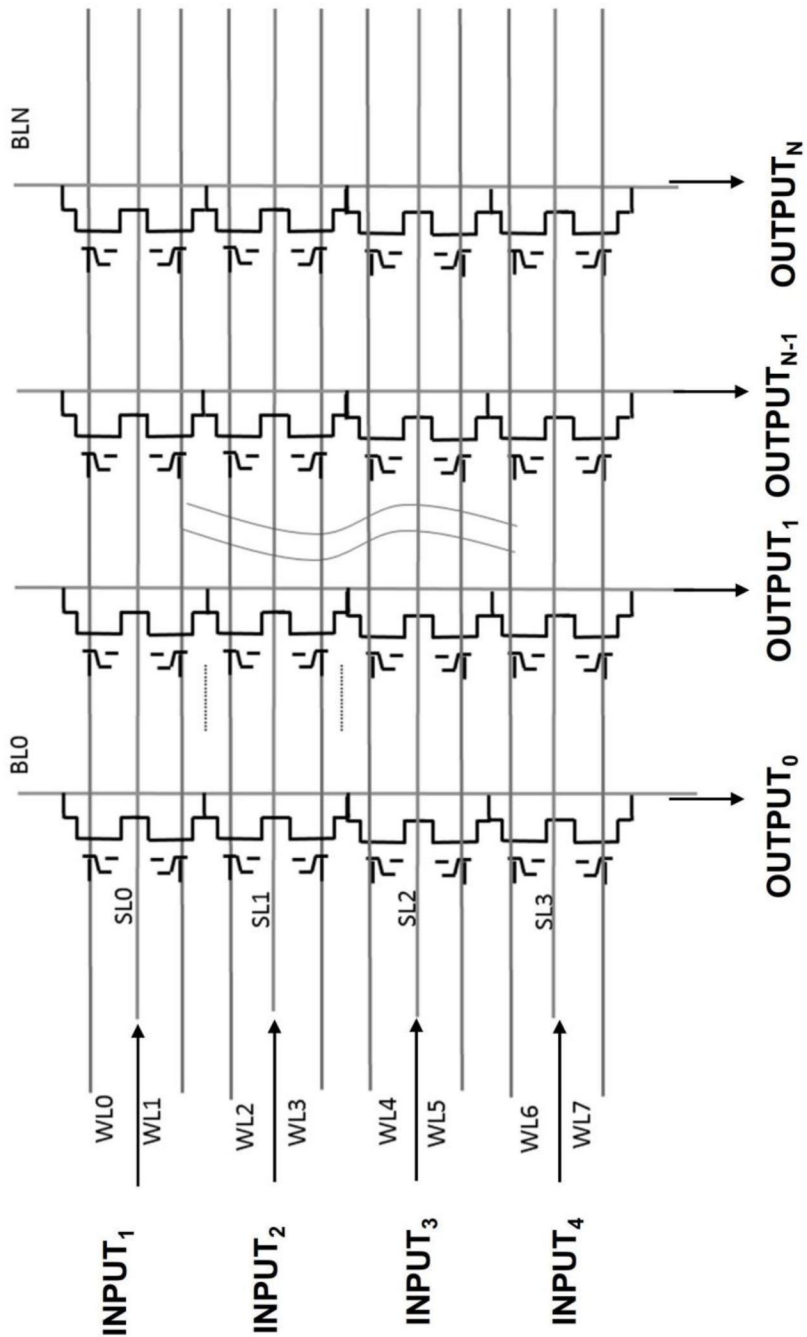


【圖21】

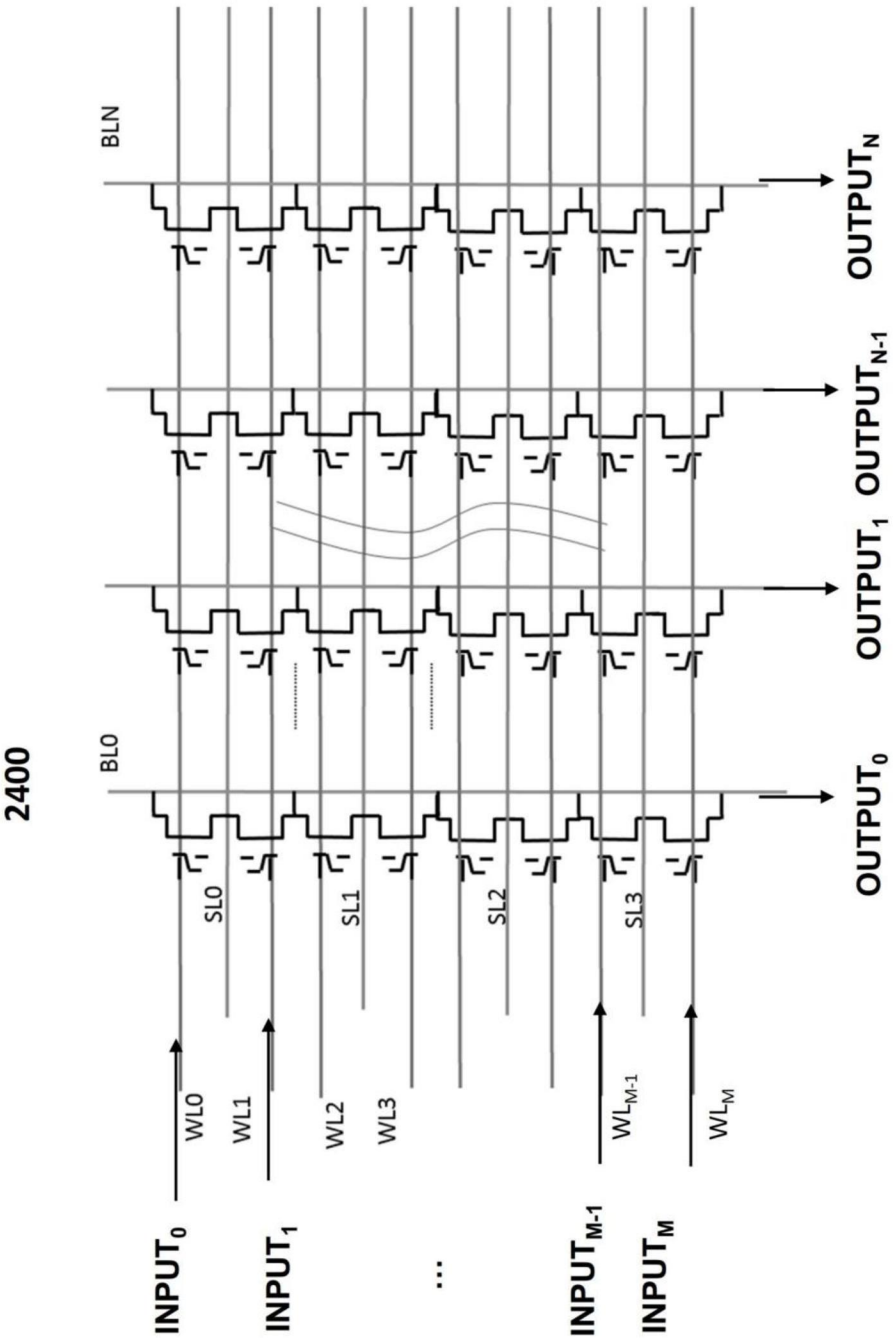


【圖22】

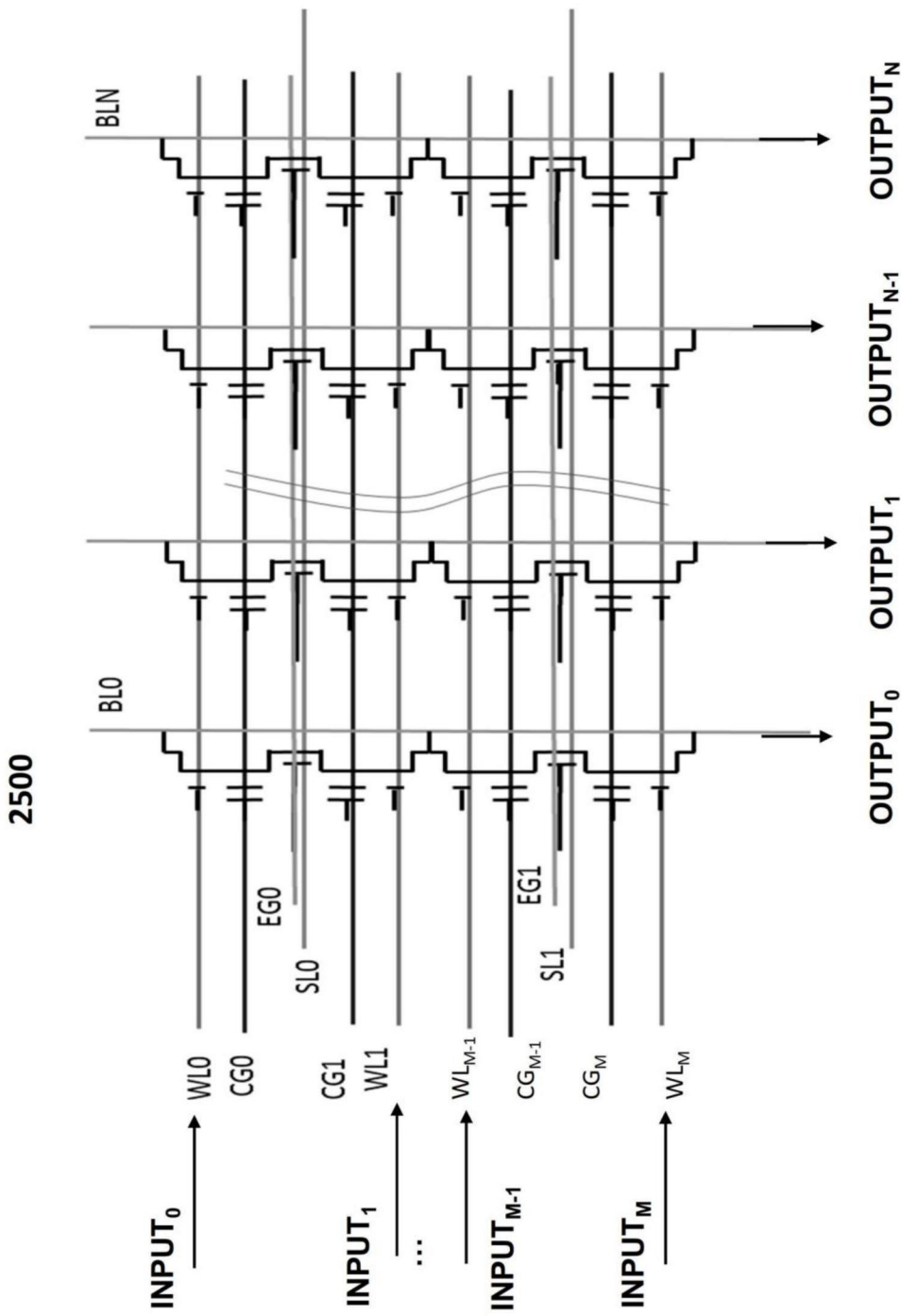
2300



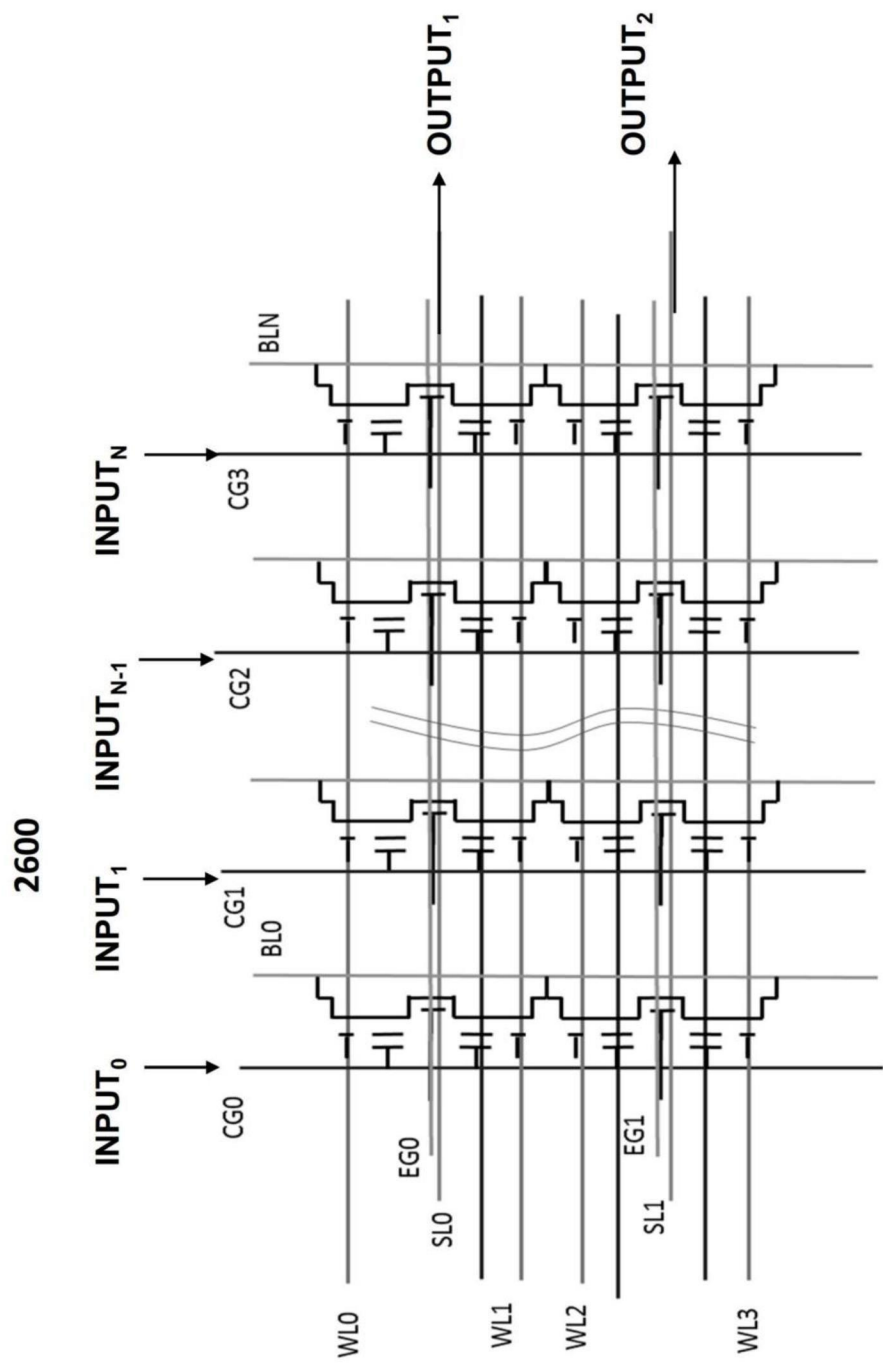
【圖23】



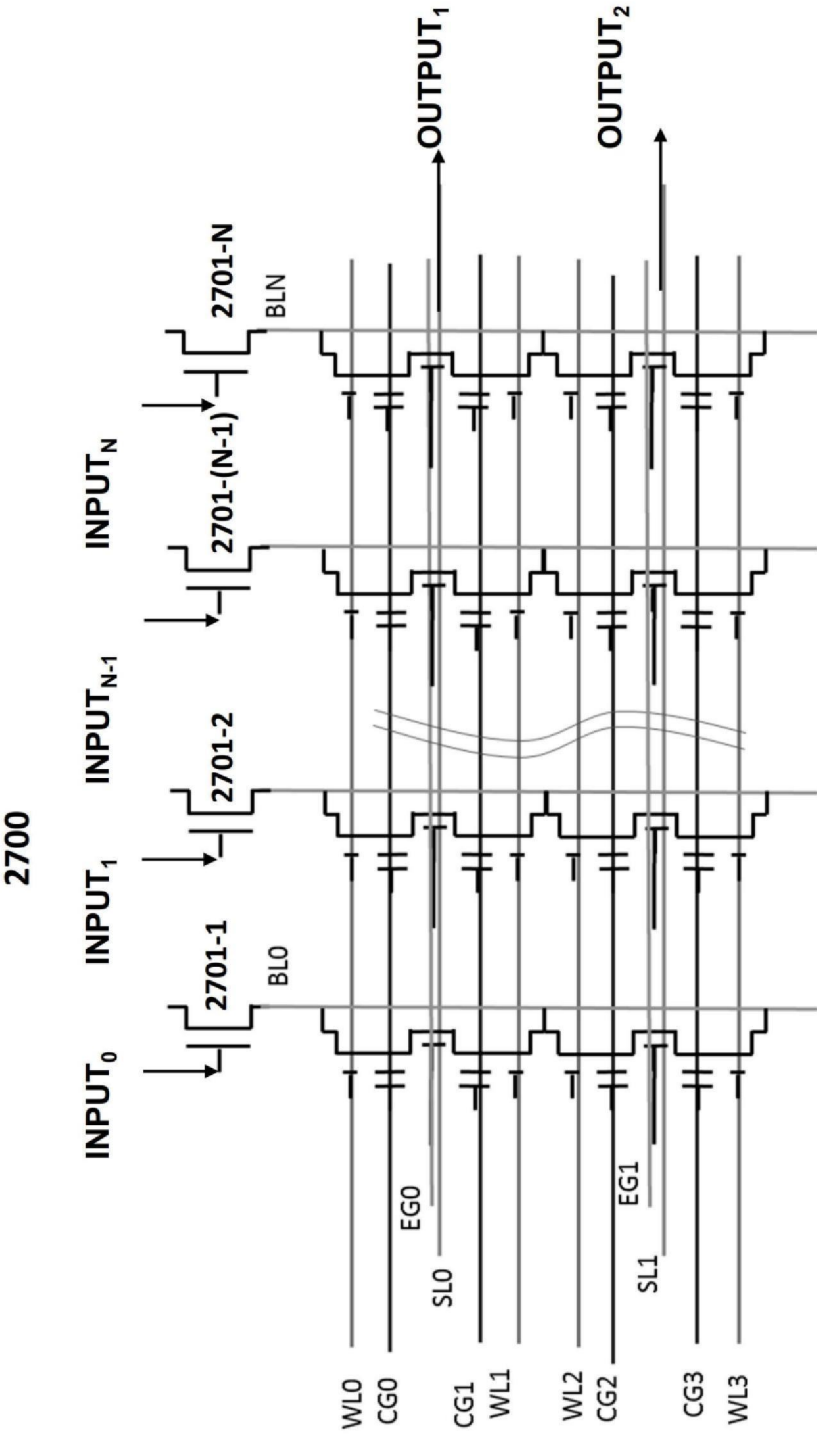
【圖24】



【圖25】

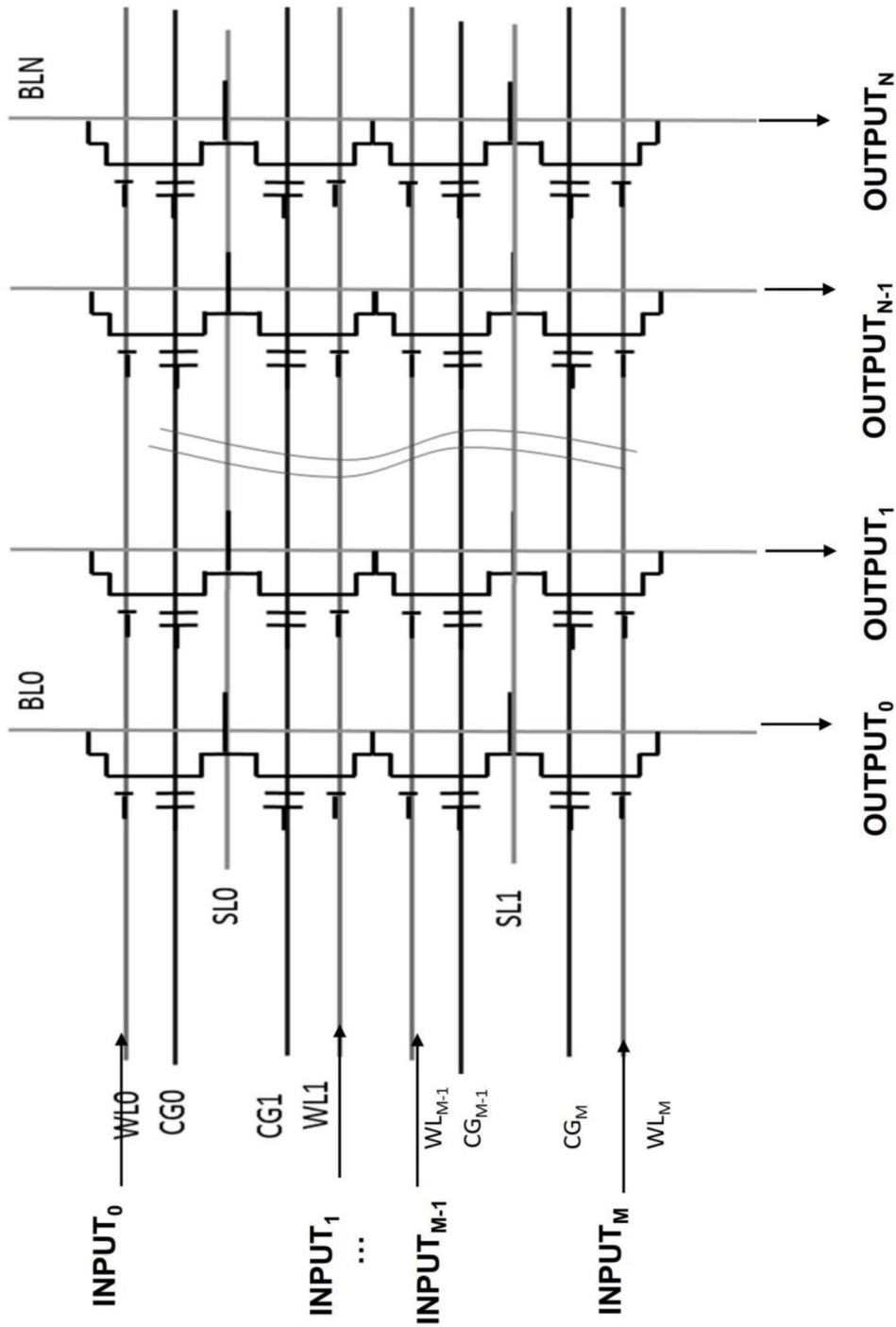


【圖26】



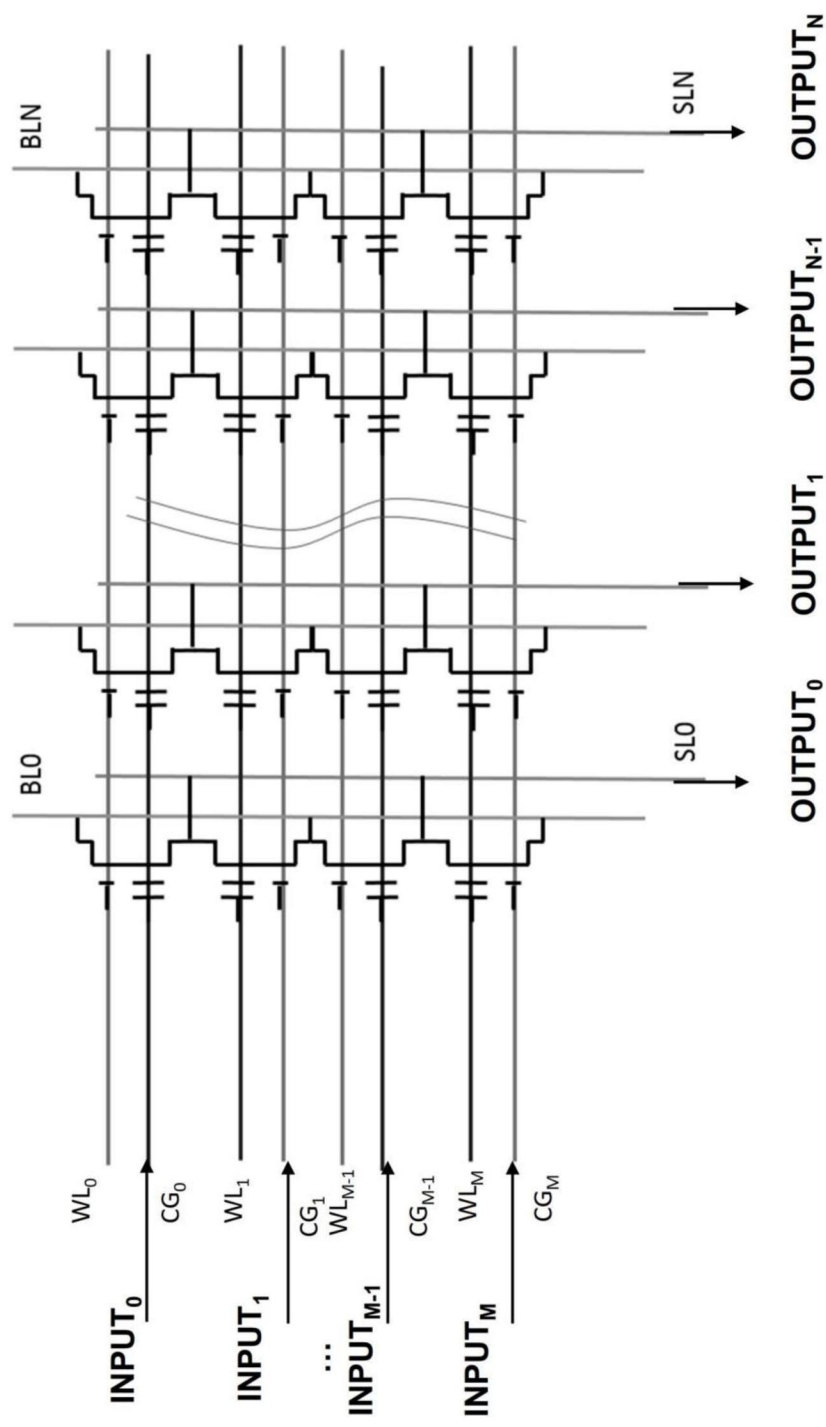
【圖27】

2800

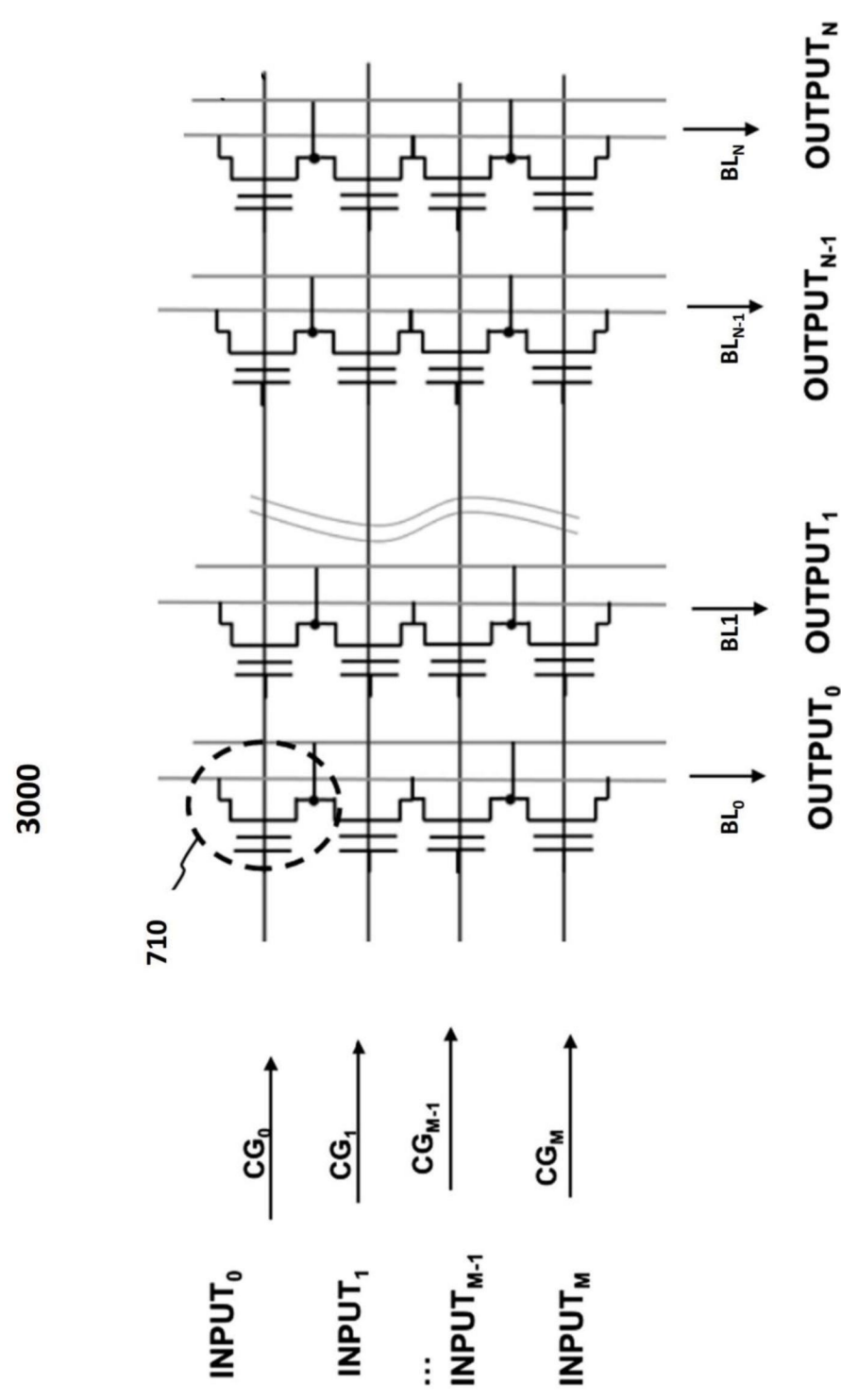


【圖28】

2900

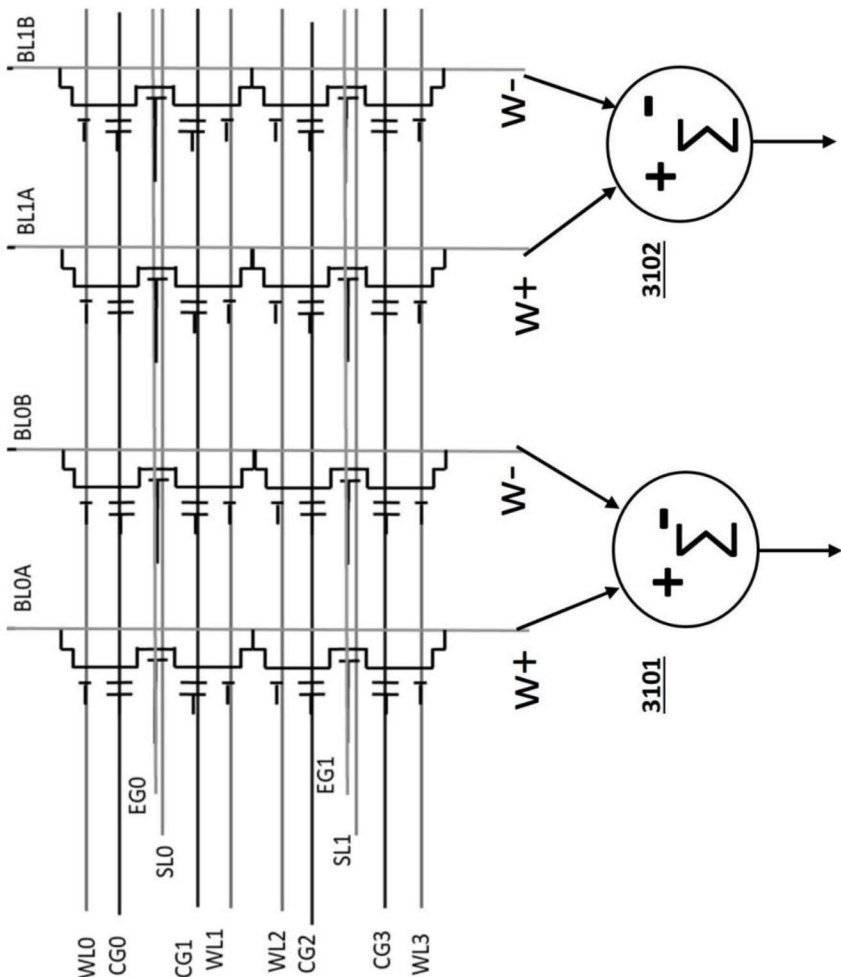


【圖29】

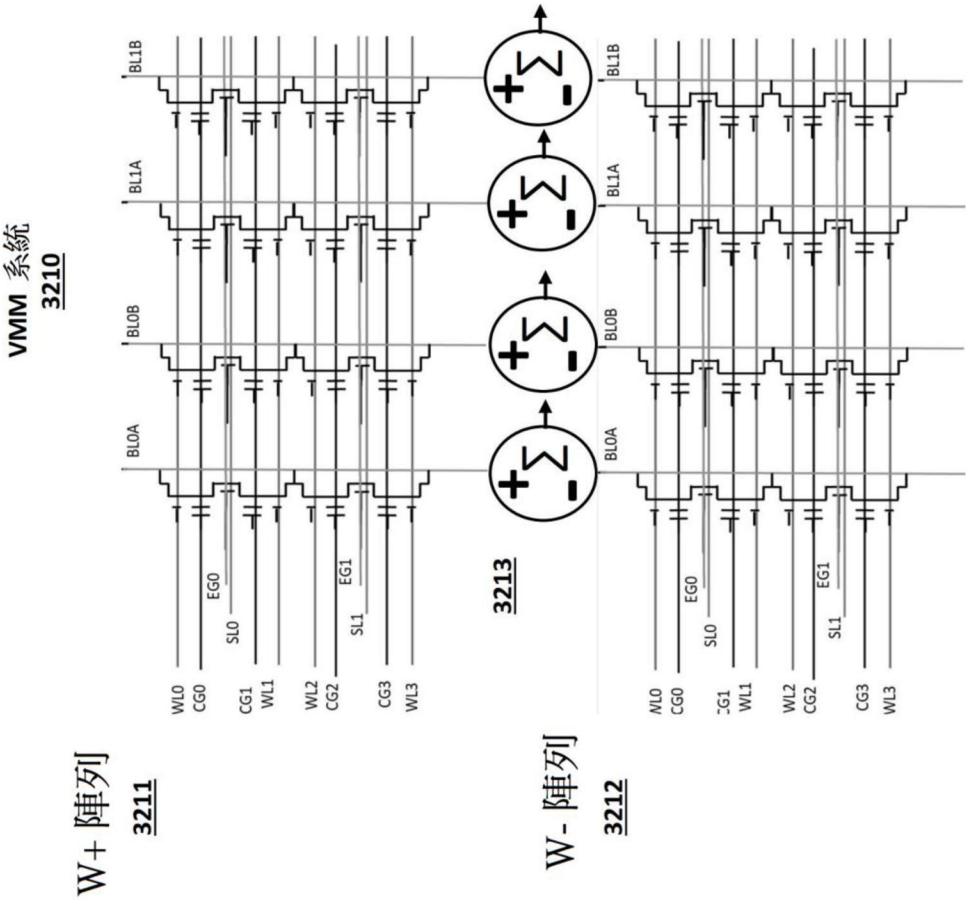


【圖30】

VMM 系統
3100

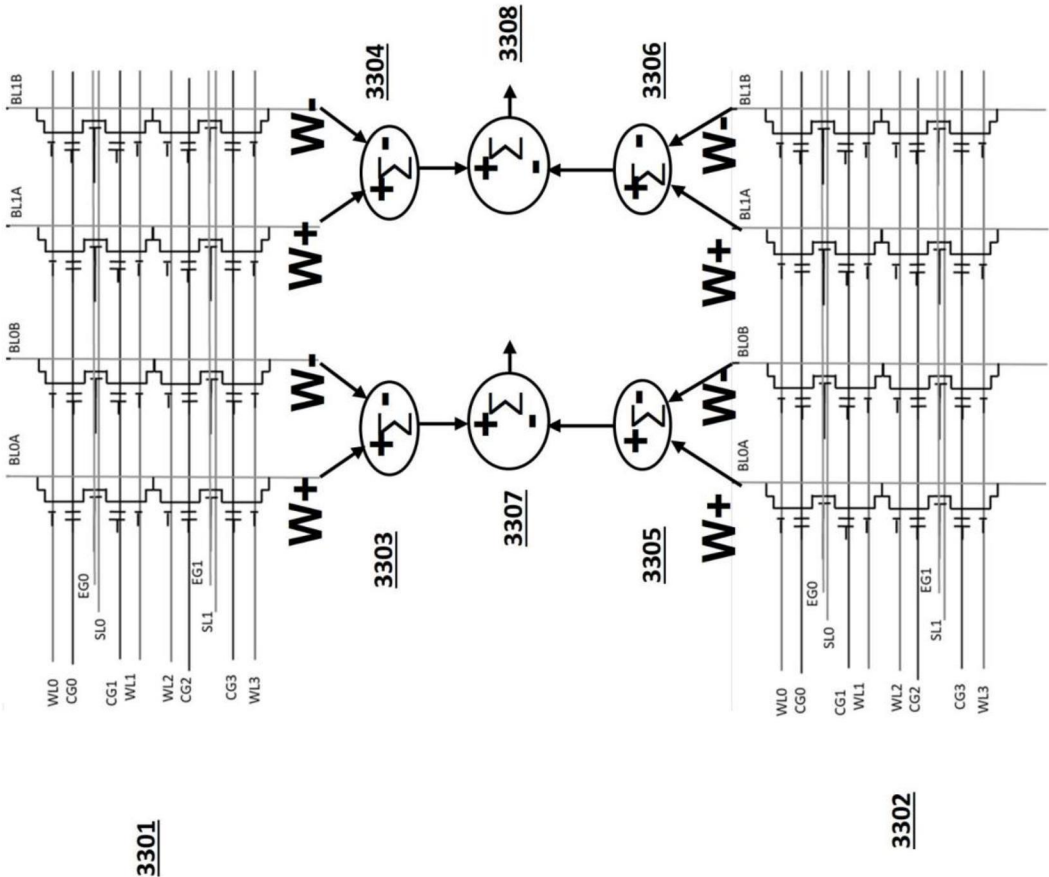


【圖31】



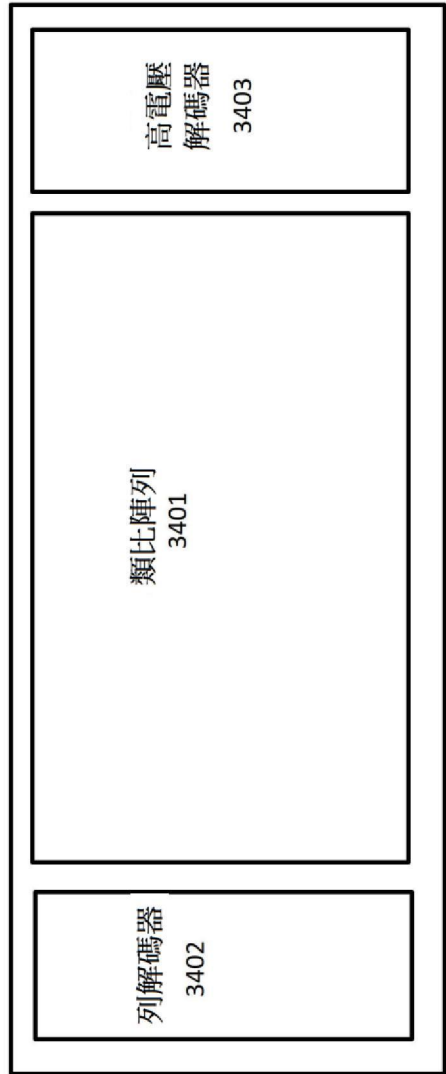
【圖32】

VMM 系統
3300



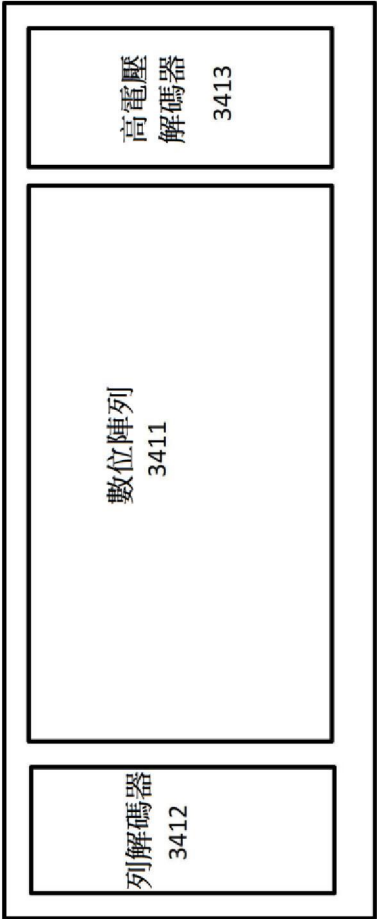
【圖33】

VMM 系統
3400

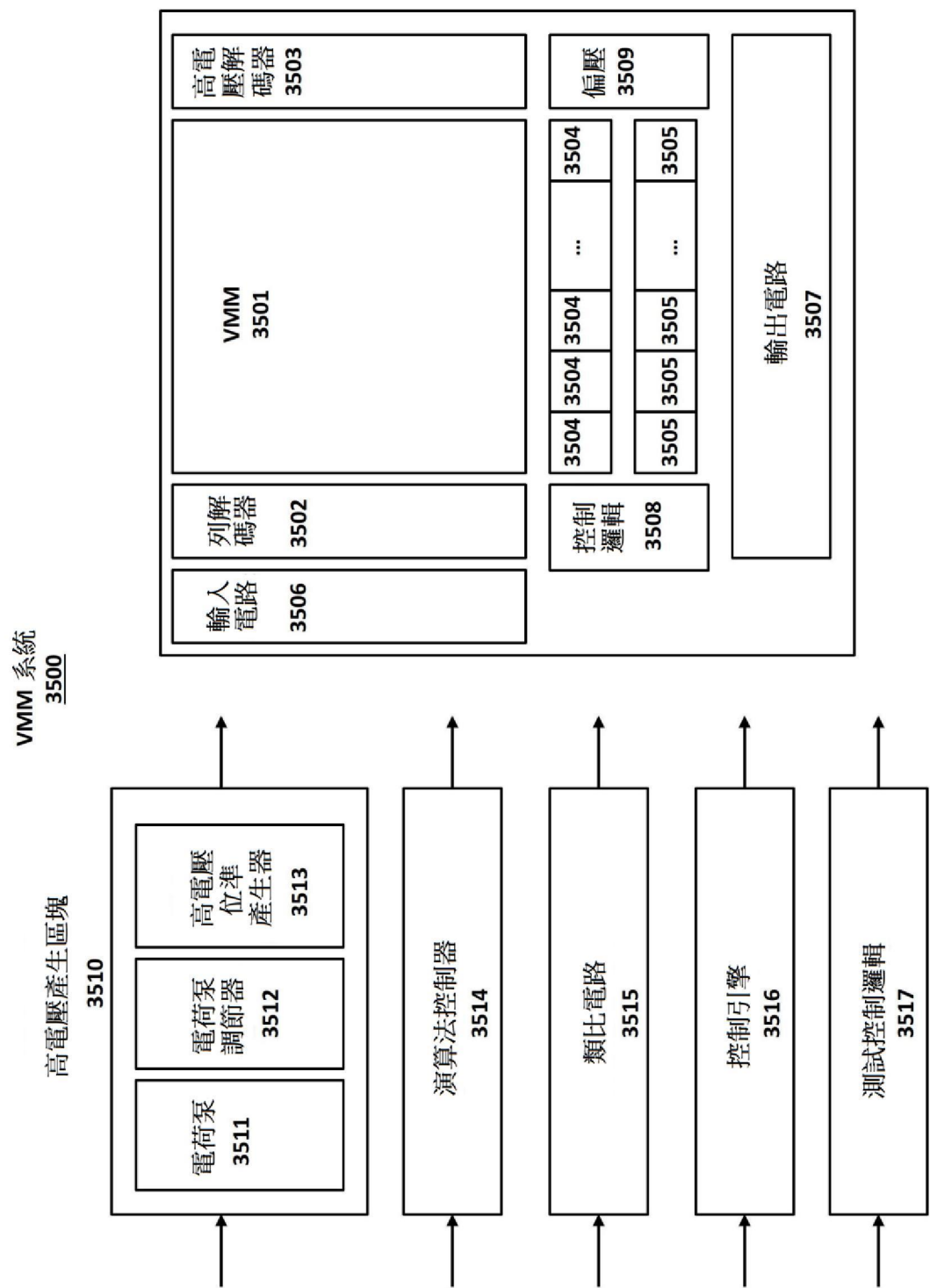


類比陣列區塊
3405

數位陣列區塊
3415

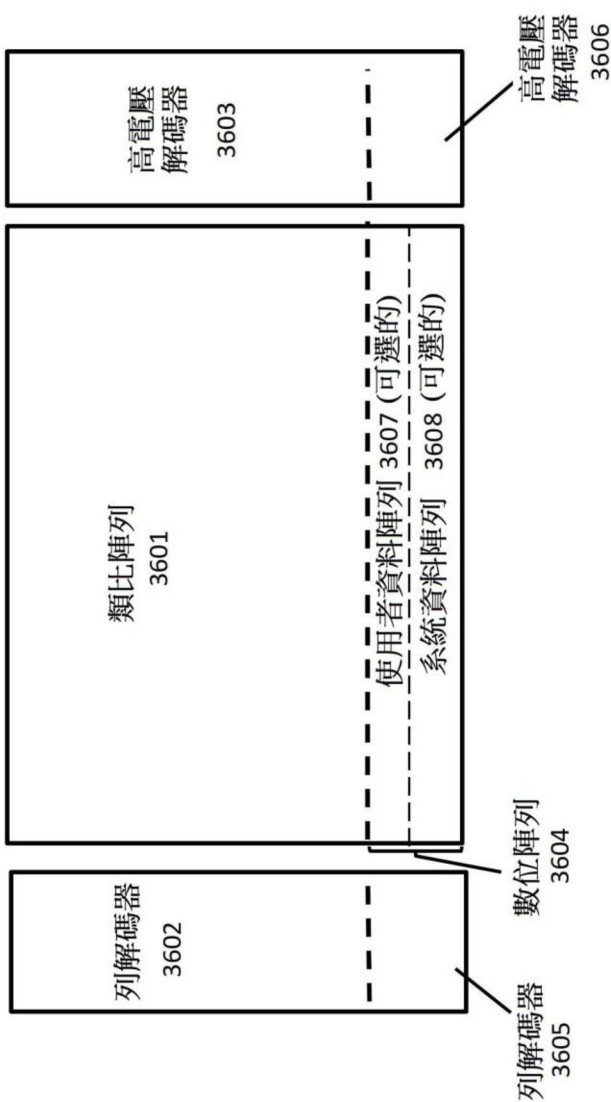


【圖34】



【圖35】

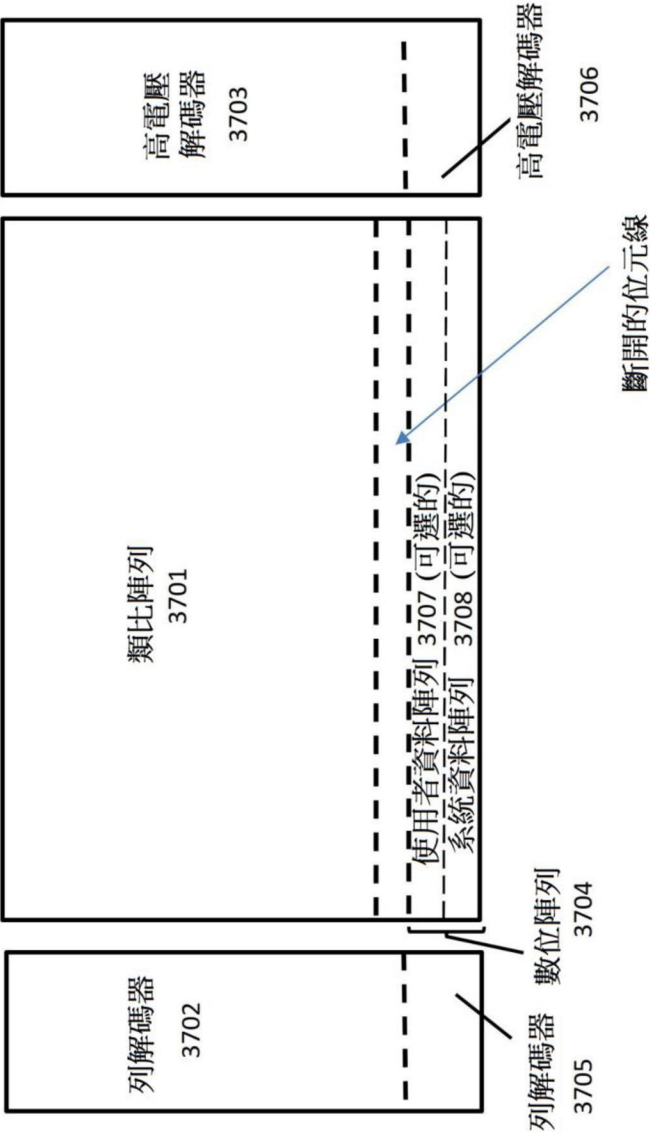
VMIM 系統
3600



可選地：
(對系統資料陣列3608的讀取、程式化、抹除操作之速度) < (對使用者資料陣列3607的讀取、程式化、抹除操作之速度)
(系統資料陣列3608的耐久性) < (使用者資料陣列3607的耐久性)

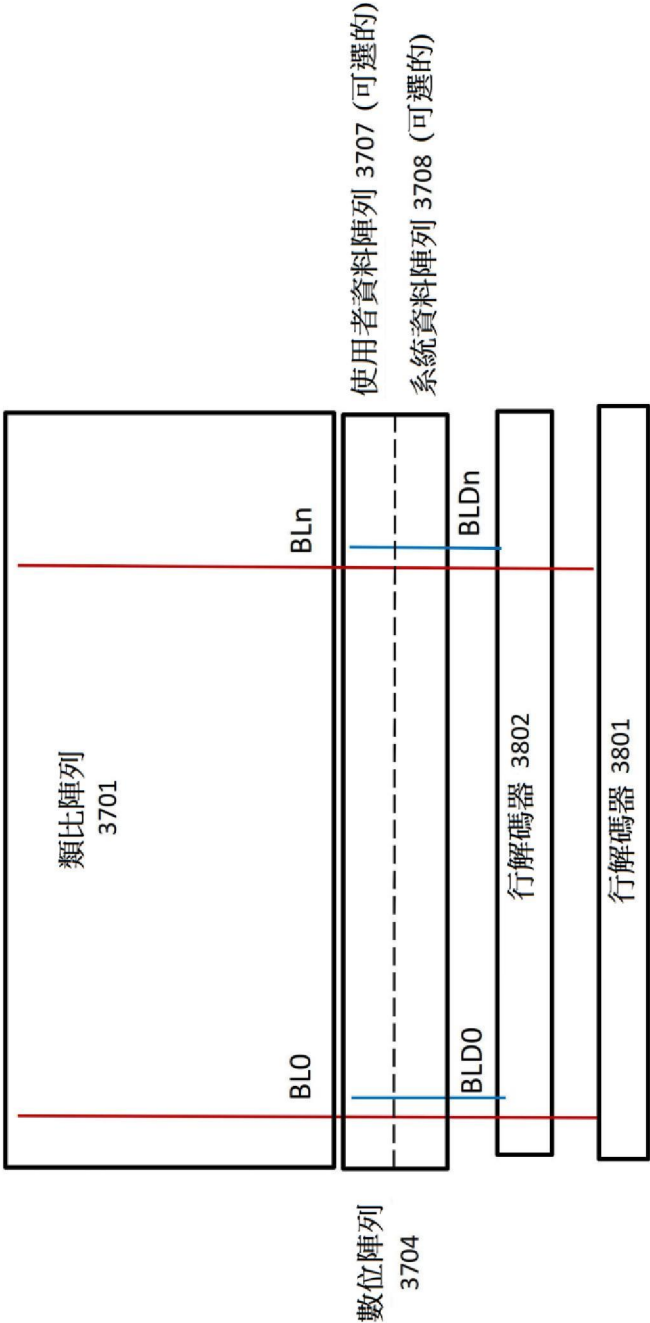
【圖36】

VMM 系統
3700

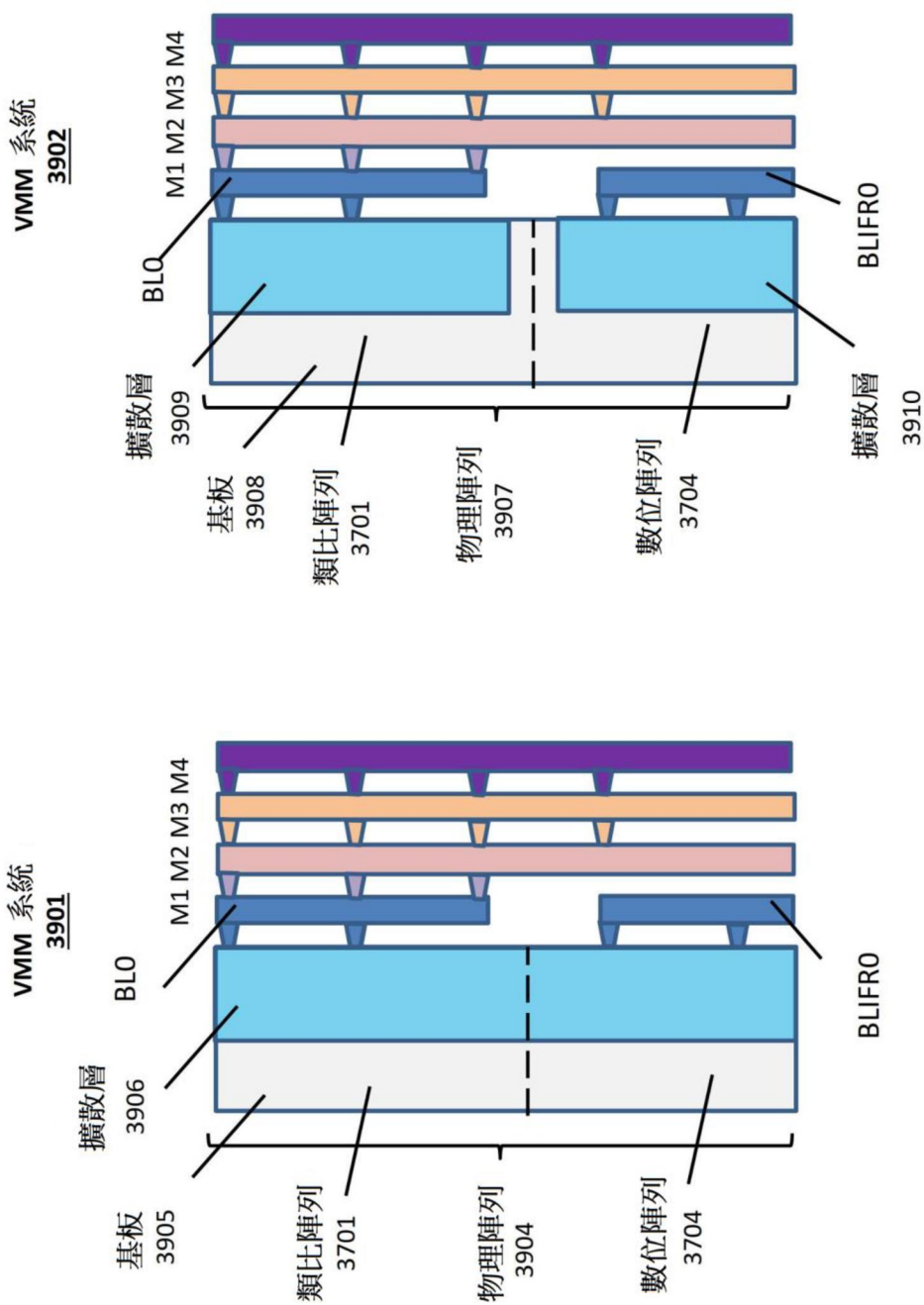


【圖37】

VMM 系統
3700

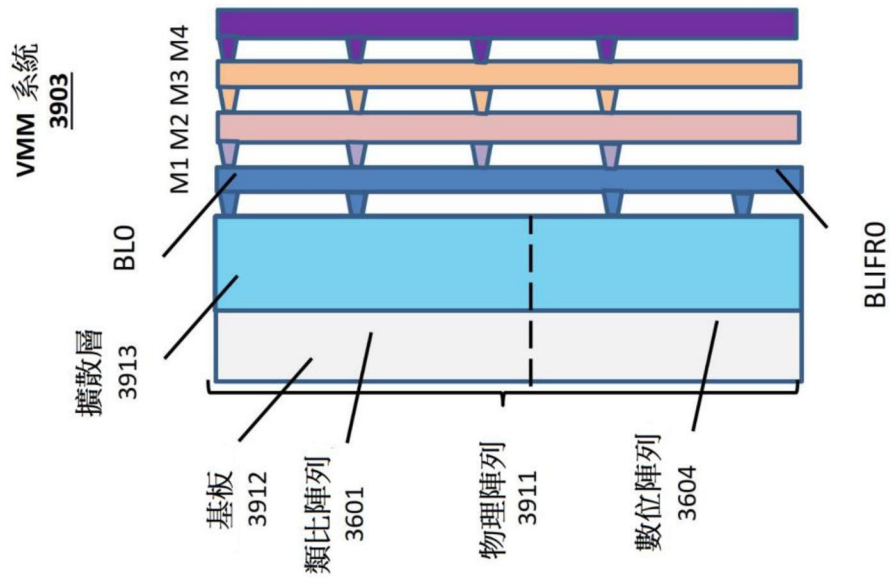


【圖38】



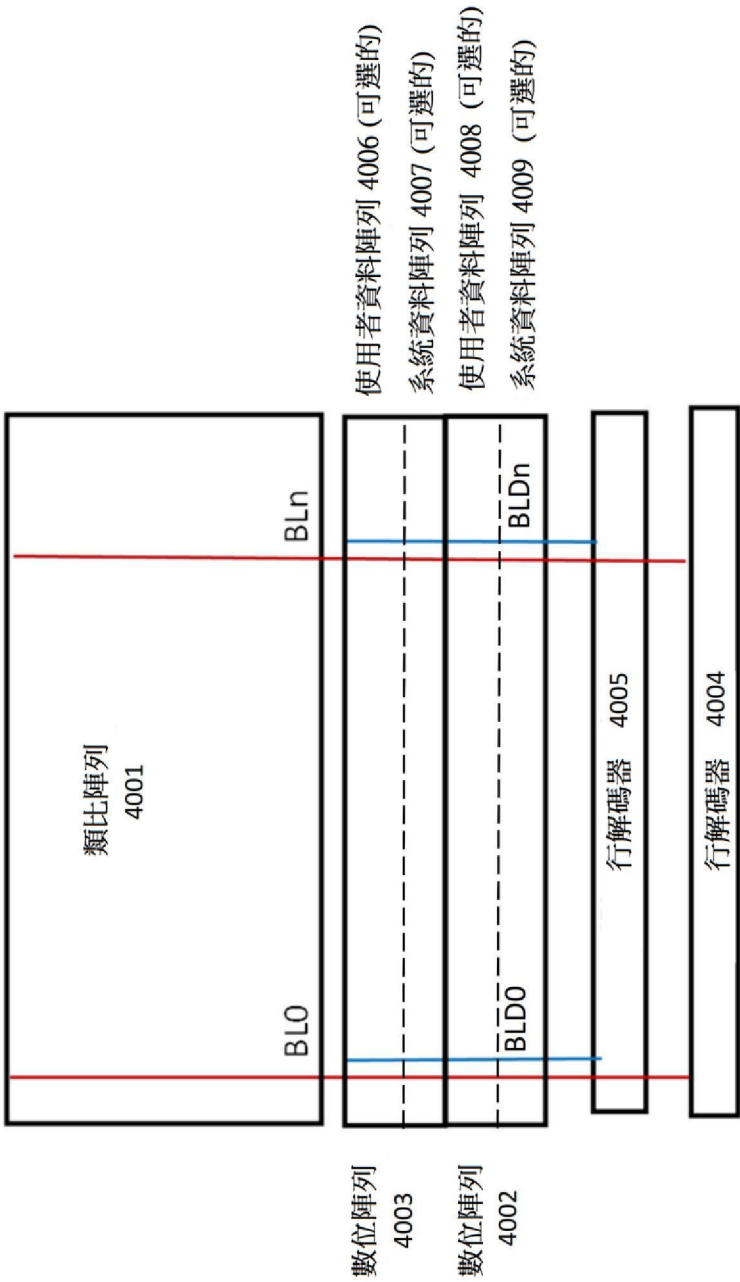
【圖39B】

【圖39A】

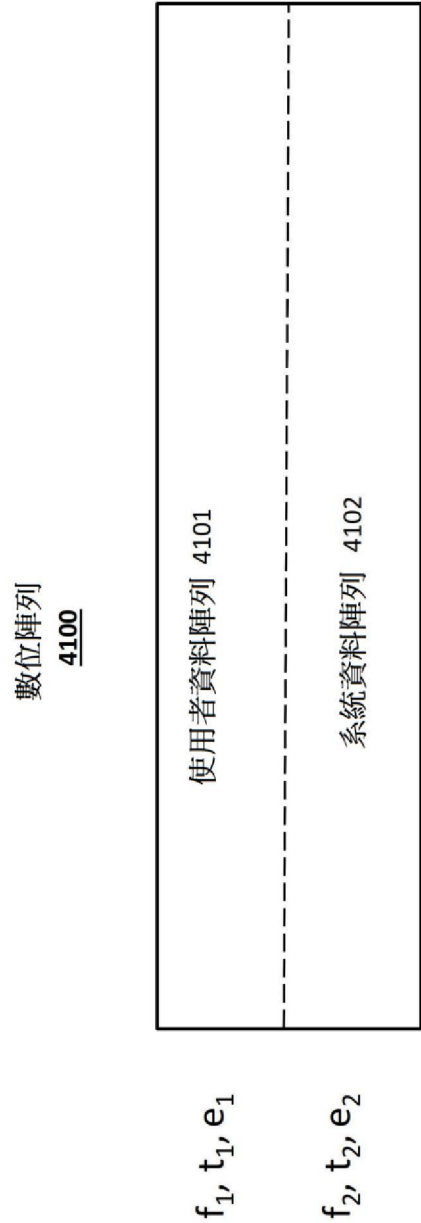


【圖39C】

VMM 系統
4000

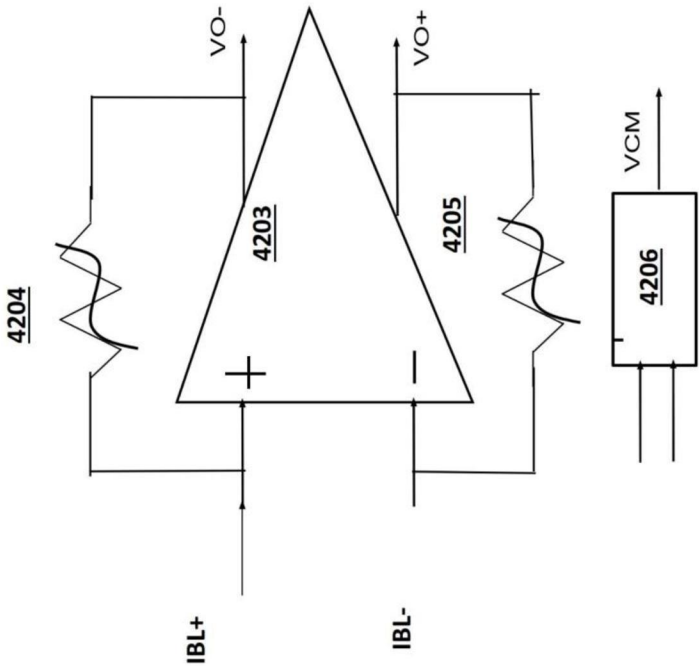


【圖40】

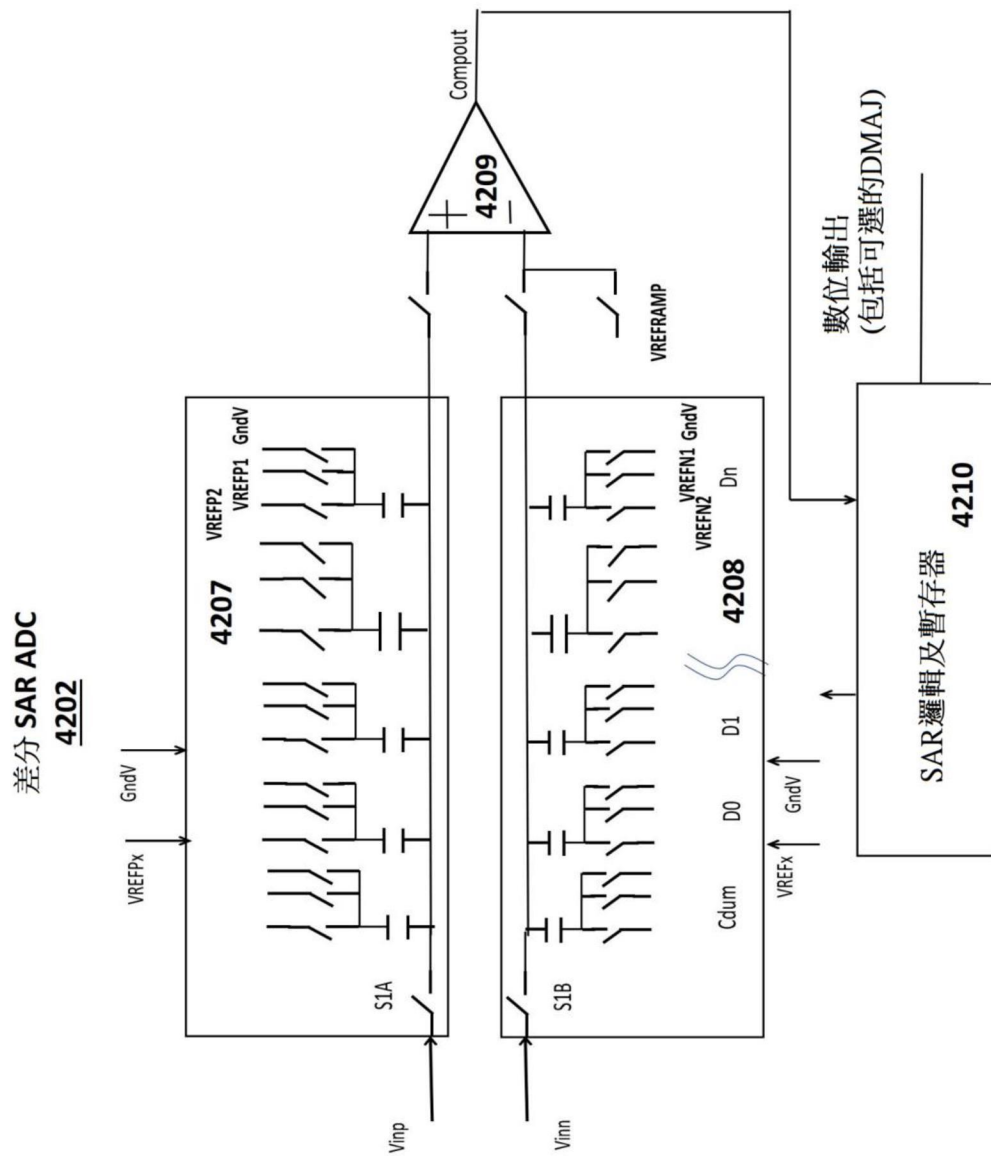


【圖41】

差分電流至
電壓轉換器
4201

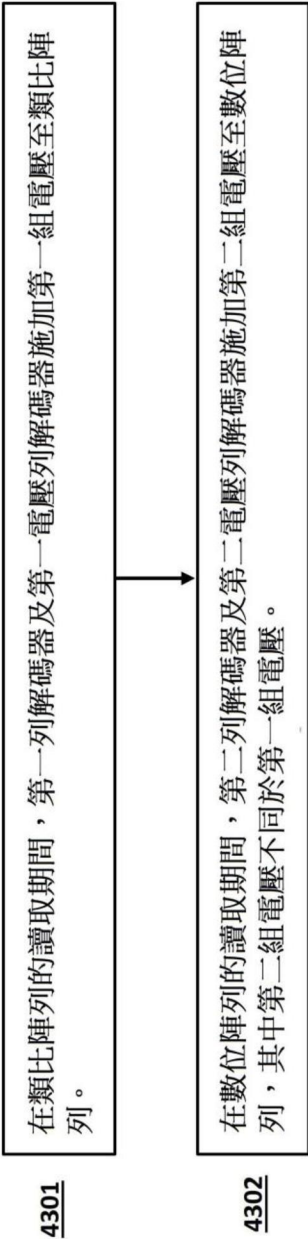


【圖42A】



【圖42B】

讀取操作
4300



【圖43】