



(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 I740487 B

(45)公告日：中華民國 110 (2021) 年 09 月 21 日

(21)申請案號：109114909

(22)申請日：中華民國 107 (2018) 年 11 月 28 日

(51)Int. Cl. : G11C16/10 (2006.01)

G11C16/34 (2006.01)

G06N3/06 (2006.01)

(30)優先權：2017/11/29 美國

15/826,345

2018/10/30 世界智慧財產權組織

PCT/US18/58297

(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)；蒂瓦里 維平 TIWARI, VIPIN (US)；杜 恩
漢 DO, NHAN (US)；利姆克 史蒂芬 LEMKE, STEVEN (US)；哈里哈蘭 聖托
西 HARIHARAN, SANTOSH (IN)；洪 史丹利 HONG, STANLEY (US)

(74)代理人：劉法正；尹重君

(56)參考文獻：

US 2005/0167729A1

US 2006/0104120A1

US 2009/0103361A1

US 2013/0044544A1

US 2014/0054667A1

US 2015/0324691A1

US 2016/0093382A1

US 2016/0180945A1

US 2016/0240255A1

US 2017/0125103A1

審查人員：蔡明宏

申請專利範圍項數：57 項 圖式數：50 共 104 頁

(54)名稱

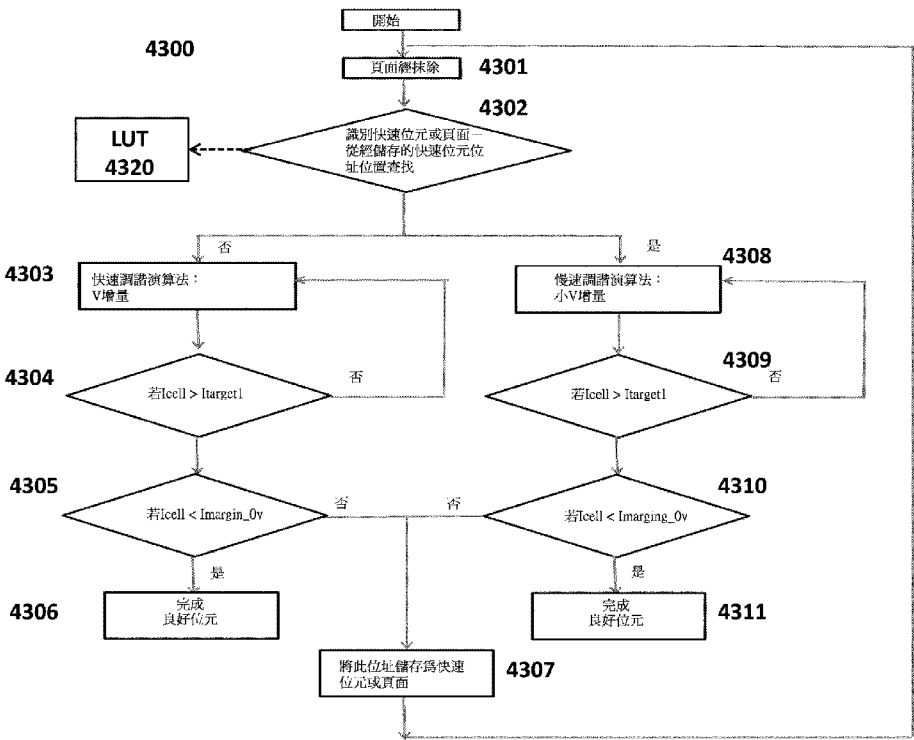
用於人工類神經網路中之類比神經形態記憶體之高精確度及高效率調諧機制及演算法

(57)摘要

一種人工類神經網路裝置，其利用包含一或多個非揮發性記憶體陣列的類比神經形態記憶體。實施例包含經改善的機制及演算法以用於調諧非揮發性記憶體陣列，使得記憶體單元的浮閘可使用所欲的電荷量迅速而精確地注入以表示由人工類神經網路用作一權重之一類比值。

An artificial neural network device that utilizes analog neuromorphic memory that comprises one or more non-volatile memory arrays. The embodiments comprise improved mechanisms and algorithms for tuning the non-volatile memory arrays such that the floating gates of the memory cells can be quickly and accurately injected with the desired amount of charge to signify an analog value utilized as a weight by the artificial neural network.

指定代表圖：



【圖43】

符號簡單說明：

4300:調諧演算法

4301:步驟

4302:步驟

4303:步驟

4304:步驟

4305:步驟

4306:步驟

4307:步驟

4308:步驟

4309:步驟

4310:步驟

4311:步驟

4320:查找表



I740487申請日: 107年11月28日【發明摘要】

IPC 分類號: G11C 16/10 (2006.01)
G11C 16/34 (2006.01)
G06N 3/06 (2006.01)

【中文發明名稱】 用於人工類神經網路中之類比神經形態記憶體
之高精度度及高效率調諧機制及演算法

【英文發明名稱】 HIGH PRECISION AND HIGHLY EFFICIENT
TUNING MECHANISMS AND ALGORITHMS FOR ANALOG
NEUROMORPHIC MEMORY IN ARTIFICIAL NEURAL
NETWORKS

【中文】

一種人工類神經網路裝置，其利用包含一或多個非揮發性記憶體陣列的類比神經形態記憶體。實施例包含經改善的機制及演算法以用於調諧非揮發性記憶體陣列，使得記憶體單元的浮閘可使用所欲的電荷量迅速而精確地注入以表示由人工類神經網路用作一權重之一類比值。

【英文】

An artificial neural network device that utilizes analog neuromorphic memory that comprises one or more non-volatile memory arrays. The embodiments comprise improved mechanisms and algorithms for tuning the non-volatile memory arrays such that the floating gates of the memory cells can be quickly and accurately injected with the desired amount of charge to signify an analog value utilized as a weight by the artificial neural network.

【指定代表圖】 第（ 43 ）圖。

【代表圖之符號簡單說明】：

4300...調諧演算法	4308...步驟
4301...步驟	4309...步驟
4302...步驟	4310...步驟
4303...步驟	4311...步驟
4304...步驟	4320...查找表
4305...步驟	4306...步驟
4307...步驟	

【特徵化學式】

(無)

【發明說明書】

【中文發明名稱】 用於人工類神經網路中之類比神經形態記憶體之高精度度及高效率調諧機制及演算法

【英文發明名稱】 HIGH PRECISION AND HIGHLY EFFICIENT TUNING MECHANISMS AND ALGORITHMS FOR ANALOG NEUROMORPHIC MEMORY IN ARTIFICIAL NEURAL NETWORKS

【技術領域】

相關申請案之交互參考

【0001】 本專利申請案主張 2017 年 11 月 29 日申請之美國專利申請案第 15/826,345 號的優先權。

【0002】 揭示針對在一人工類神經網路中使用之一類比神經形態記憶體內之調諧單元的眾多實施例。

【先前技術】

【0003】 人工類神經網路模仿生物神經網路（動物的中樞神經系統，特別係大腦），其用以評估或近似可取決於大數目的輸入且通常未知的功能。人工類神經網路通常包括在彼此間交換訊息之互連的「神經元」的層。圖 1 繪示一人工類神經網路，其中圓形表示神經元的輸入或層。連接（稱為突觸）係以箭頭表示，並具有可基於經驗調諧的數字權重。此使類神經網路適應輸入並能夠學習。一般而言，類神經網路包括多個輸入的一層。一般而言有一或多個神經元中間層，及提供類神經網路的輸出之一神經元輸出層。各級的神經元基於自突觸接收之資料個別地或共同地做決策。

【0004】 用於高效能資訊處理的人工類神經網路之發展中的主要挑戰之一者係缺少適當的硬體技術。的確，實際的類神經網路依賴非常大數目的突觸，而在神經元之間實現高連接性，亦即，非常高的計算平行性。原理上，此種複雜性可以數位超級電腦或專用圖形處理單元叢集實現。然而，除了高成本以外，此等方法與生物網路相比時也苦於中等的能量效率，主要因為生物網路實施低精確度類比計算而消耗更少的能量。CMOS 類比電路已用於人工類神經網路，但大多數 CMOS 實作的突觸在給定高數目的神經元及突觸情況下已過於龐大。

【0005】 申請人先前在美國專利申請案第 15/594,439 號（其以引用方式併入本文中）中揭示一種人工（類比）類神經網路，其利用一或多個非揮發性記憶體陣列作為突觸。非揮發性記憶體陣列作為類比神經形態記憶體進行操作。該類神經網路裝置包括一第一複數個突觸及一第一複數個神經元，該第一複數個突觸經組態以接收一第一複數個輸入並由此產生一第一複數個輸出，該第一複數個神經元經組態以接收該第一複數個輸出。該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之各者包括形成於一半導體基材中之間隔開的源極區域及汲極區域，其中一通道區域在該源極區域與該汲極區域之間延伸；一浮閘，其設置於該通道區域之一第一部分上方並與該第一部分絕緣；及一非浮閘，其設置於該通道區域之一第二部分上方並與該第二部分絕緣。該複數個記憶體單元之各者經組態來儲存對應於該浮閘上之數個電子的權重值。該複數個記憶體單元經組態以使該第一複數個輸入乘以所儲存的該等權重值，以產生該第一複數個輸出。

【0006】 該類比神經形態中所用的各非揮發性記憶體單元必須經抹除及程式化以在該浮閘中保持一非常特定且精確的電荷量。例如，各浮閘必須保持 N 個不同值之一者，其中 N 係可由各單元指示之不同權重的數目。 N 的實例包括 16、32、及 64。先前技術缺乏用於調諧各單元之一快速且精確的機制以確保該單元含有所欲的電荷量。

【0007】 所需的係經改善的機制及演算法，以用於調諧人工類神經網路中所用之一類比神經形態記憶體。

【發明內容】

【0008】 於本揭示的一個態樣中，本發明揭露一種用於在非揮發性記憶體單元之一陣列內識別可經迅速程式化的非揮發性記憶體單元之方法，該方法包含抹除一非揮發性記憶體單元；以一第一電壓程式化該非揮發性記憶體單元；測量在一讀取操作期間通過該非揮發性記憶體單元之一第一電流；以一第二電壓程式化該非揮發性記憶體單元；測量在一讀取操作期間通過該非揮發性記憶體單元之一第二電流；及若該第二電流與該第一電流之間的差異超過一臨限值，儲存指示該非揮發性記憶體單元可經迅速程式化的資料。

【圖式簡單說明】

【0009】

圖 1 係繪示一人工類神經網路的圖。

圖 2 係習知的 2 閘極非揮發性記憶體單元的側截面圖。

圖 3 係繪示用於圖 2 之記憶體單元之一習知陣列架構的圖。

圖 4 係習知的 2 閘極非揮發性記憶體單元的側截面圖。

圖 5 係繪示用於圖 4 之記憶體單元之一習知陣列架構的圖。

圖 6 係習知的 4 閘極非揮發性記憶體單元的側截面圖。

圖 7 係繪示用於圖 6 之記憶體單元之一習知陣列架構的圖。

圖 8A 係繪示均勻間隔之類神經網路權重等級分派的圖。

圖 8B 係繪示不均勻間隔之類神經網路權重等級分派的圖。

圖 9 係繪示一雙向調諧演算法的流程圖。

圖 10 係繪示使用電流比較之權重映射的方塊圖。

圖 11 係繪示使用電壓比較之權重映射的方塊圖。

圖 12 係繪示利用一非揮發性記憶體陣列之一例示性類神經網路之不同等級的圖。

圖 13 係繪示一向量乘法器矩陣的方塊圖。

圖 14 係繪示一向量乘法器矩陣之各種級的方塊圖。

圖 15 至圖 16 係繪示四閘極記憶體單元的陣列之第一架構的示意圖。

圖 17 至圖 18 係繪示四閘極記憶體單元的陣列之第二架構的示意圖。

圖 19 係繪示四閘極記憶體單元之一陣列之一第三架構的示意圖。

圖 20 係繪示四閘極記憶體單元之一陣列之一第四架構的示意圖。

圖 21 係繪示四閘極記憶體單元之一陣列之一第五架構的示意圖。

圖 22 係繪示四閘極記憶體單元之一陣列之一第六架構的示意圖。

圖 23 係繪示二閘極記憶體單元之一陣列之一第一架構的示意圖。

圖 24 係繪示二閘極記憶體單元之一陣列之一第二架構的示意圖。

圖 25 係繪示一電流至電壓對數轉換器的圖。

圖 26 係繪示一電壓至電流對數轉換器的圖。

圖 27 係繪示一參考 Gnd 的電流加算器的圖。

圖 28 係繪示一參考 Vdd 的電流加算器的圖。

圖 29 係繪示一非揮發性記憶體陣列之 N^2 個類神經網輸入之使用的圖。

圖 30 係繪示一非揮發性記憶體陣列之 N^2 個類神經網輸入之使用的圖。

圖 31 係繪示具有週期性移位輸入線之一非揮發性記憶體陣列之類神經網輸入之使用的圖。

圖 32 係繪示圖 15 之記憶體陣列架構的示意圖，但具有週期性移位輸入線。

圖 33 係繪示圖 20 之記憶體陣列架構的示意圖，但具有週期性移位輸入線。

圖 34 係可用於一人工類神經網路中之一先前技術非揮發性記憶體單元的截面圖。

圖 35 描繪非揮發性記憶體單元的一陣列，其用於在一人工類神經網路中使用。

圖 36 描繪用以在圖 35 之非揮發性記憶體單元上執行操作之操作電壓。

圖 37 描繪非揮發性記憶體單元的一陣列，其用於在一人工類神經網路中使用。

圖 38 描繪用以在圖 37 之非揮發性記憶體單元上執行操作之操作電壓。

圖 39 描繪非揮發性記憶體單元的一陣列，其用於在一人工類神經網路中使用。

圖 40 描繪用以在圖 39 之非揮發性記憶體單元上執行操作之操作電壓。

圖 41 描繪非揮發性記憶體單元的一陣列，其用於在一人工類神經網路中使用。

圖 42 描繪用以在圖 41 之非揮發性記憶體單元上執行操作之操作電壓。

圖 43 描繪一演算法，其用於在非揮發性記憶體單元之一陣列內以不同的程式化速度特性調諧單元。

圖 44 描繪一演算法，其用於識別可在非揮發性記憶體單元之一陣列內經迅速程式化的單元。

圖 45 描繪一演算法，其用於調諧包含一粗略程式化序列及一精細程式化序列之單元。

圖 46 描繪一均勻步階演算法，其用於調諧一非揮發性記憶體單元。

圖 47 描繪一均勻對數步階演算法，其包含用於調諧一非揮發性記憶體單元的粗略及精細步階。

圖 48 描繪一二元搜尋步階演算法，其包含用於調諧一非揮發性記憶體單元的粗略及精細步階。

圖 49 描繪用於一非揮發性記憶體單元之一調諧序列的一波形。

圖 50 描繪用於一非揮發性記憶體單元之一調諧序列的一波形。

【實施方式】

【0010】 本發明的人工類神經網路利用 CMOS 技術與非揮發性記憶體陣列的組合。數位非揮發性記憶體係已知。例如，美國專利第 5,029,130 號（「'130 專利」）揭示分離閘非揮發性記憶體單元之一陣列，且係為所有目的以引用方式併入本文中。該記憶體單元係顯示於圖 2 中。各記憶體單元 10 包括形成於一半導體基材 12 中之源極區域 14 與汲極區域 16，其中有一通道區域 18 位於其間。一浮閘 20 形成於通道區域 18 之一第一部分上方且與該第一部分絕緣（且控制該第一部分的導電性），並形成於汲極區域 16 的一部分上方。一控制閘 22 具有一第一部分 22a 及一第二部分 22b，第一部分 22a 設置於通道區 18 之一第二部分上方且與該第二部分絕緣（且控制該第二部分的導電性），第二部分 22b 向上延伸且延伸於浮閘 20 上方。浮閘 20 及控制閘 22 係藉由一閘極氧化物 26 而與基材 12 絕緣。

【0011】 藉由將一高正電壓置於控制閘 22 上來抹除記憶體單元（其中將電子自浮閘移除），其導致浮閘 20 上的電子藉由富爾諾罕穿隧自浮閘 20 穿隧通過中間絕緣體 24 至控制閘 22。

【0012】 藉由將一正電壓置於控制閘 22 上以及將一正電壓置於汲極 16 上來程式化記憶體單元（其中將電子置於浮閘上）。電子流將自源極 14 朝向汲極 16 流動。當電子抵達控制閘 22 與浮閘 20 之間的時間隙時，電子將加速且變成經加熱的。由於來自浮閘 20 的吸引靜電力，該等經加熱電子的一些將通過閘極氧化物 26 注入至浮閘 20 上。

【0013】 藉由將正讀取電壓置於汲極 16 及控制閘 22 上來讀取記憶體單元（其接通控制閘下方的通道區域）。若浮閘 20 帶正電荷（亦即電子經抹除並正耦接至汲極 16），則浮閘 20 下方的通道區域部分亦經接通，且電流將跨通道區域 18 流動，其係感測為經抹除或「1」狀態。若浮閘 20 帶負電荷（亦即以電子程式化），則浮閘 20 下方的通道區域部分係大部分或完全斷開，且電流將不會跨通道區域 18 流動（或將有少許流動），其係感測為經程式化或「0」狀態。

【0014】 先前技術記憶體陣列的架構係顯示於圖 3 中。記憶體單元 10 經以列與行配置。在各行中，記憶體單元以鏡像方式端對端地配置，以使其等形成為成對的記憶體單元，各對共享一共用的源極區域 14(S)，且各組相鄰的記憶體單元對共享一共用的汲極區域 16(D)。用於任何給定之記憶體單元列的所有源極區域 14 係藉由一源極線 14a 電氣連接在一起。用於任何給定之記憶體單元行的所有汲極區域 16 係藉由一位元線 16a 電氣連接在一起。用於任何給定之記憶體單元列的所有控制閘 22 係藉由一控制閘線 22a 電氣連接在一起。因此，雖然記憶體單元可經個別地程式化及讀取，但記憶體單元之抹除係一系列地執行（各列記憶體單元係藉由施加一高電壓在控制閘線 22a 上而一起抹除）。若欲抹除一特定記憶體單元，則也將相同列中的所有記憶體單元抹除。

【0015】 所屬技術領域中具有通常知識者瞭解源極與汲極可互換，其中浮閘可於源極而非汲極上方部分地延伸，如圖 4 所示（二閘極記憶體單元）。圖 5 最佳地繪示對應的記憶體單元架構，其包括記憶體

單元 10、源極線 14a、位元線 16a、及控制閘線 22a。如由圖式明顯可見者，相同列的記憶體單元 10 共享相同的源極線 14a 及相同的控制閘線 22a，而相同行之所有單元的汲極電氣連接至相同的位元線 16a。該陣列設計係針對數位應用而最佳化，並例如藉由分別施加 1.6 V 及 7.6 V 至經選擇的控制閘線 22a 及源極線 14a 並將經選擇的位元線 16a 接地，來允許經選擇單元之個別程式化。藉由在未經選取的位元線 16a 上施加一大於 2 伏的電壓並將其餘的線接地來避免干擾相同對中之非選取的記憶體單元。記憶體單元 10 無法個別地抹除，因為負責抹除的程序（電子自浮閘 20 至控制閘 22 之 Fowler-Nordheim 穿隧）僅受到汲極電壓（亦即，對在列方向中共享相同源極線 14a 的兩個相鄰單元而言唯一可係不同的電壓）的微弱影響。

【0016】具有多於兩個閘極之分離閘記憶體單元亦為已知。例如，四閘極記憶體單元具有源極區域 14、汲極區域 16、在通道區域 18 之一第一部分上方的浮閘 20、在通道區域 18 之一第二部分上方的選擇閘 28、在浮閘 20 上方的控制閘 22、及在源極區域 14 上方的抹除閘 30 係已知，如圖 6 所示（參見例如美國專利第 6,747,310 號），其係為所有目的以引用方式併入本文中。此處，除了浮閘 20 以外的所有閘係非浮閘，意指其等經電氣連接或可連接至電壓源。程式化係藉由變熱的電子自通道區域 18 將其本身注入至浮閘 20 上來顯示。抹除係藉由自浮閘 20 至抹除閘 30 之電子穿隧來顯示。

【0017】 另一類型之先前技術的分離閘三閘極記憶體單元係顯示於圖 34 中。除了不具有一分開的控制閘以外，圖 34 的分離閘記憶體

單元與圖 6 的分離閘記憶體單元完全相同。除了沒有控制閘偏壓以外，抹除操作（透過抹除閘進行抹除）與讀取操作類似於圖 6 所示者。程式化操作也並未使用控制閘偏壓，因此為了彌補對於控制閘偏壓的缺乏，源極線上的程式化電壓比較高。

【0018】 表 1 描繪執行讀取、抹除與程式化操作時，可施加至四個端子的典型電壓範圍：

表 1：圖 34 之快閃記憶體單元的操作

	WL	WL-非選取	BL	BL-非選取	EG	EG-非選取	SL	SL-非選取
讀取	0.7 至 2.2 V	-0.5 V/0 V	0.6 至 2 V	0 V/FLT	0 至 2.6 V	0 至 2.6 V	0 V	0 V/FLT/VB
抹除	-0.5 V/0 V	-.5 V/0 V	0 V	0 V	11.5 V	0 至 2.6 V	0 V	0 V
程式化	1 至 1.5 V	-.5 V/0 V	1 至 3 μ A	V _{inh} (~1.8 V)	4.5 V	0 至 2.6 V	7 至 9 V	0 至 1 V/FLT

【0019】 用於一四閘極記憶體單元陣列的架構可依圖 7 所示者來組態。在此實施例中，各水平選擇閘線 28a 將用於彼列記憶體單元的所有選擇閘 28 電氣連接在一起。各水平控制閘線 22a 將用於彼列記憶體單元的所有控制閘 22 電氣連接在一起。各水平源極線 14a 將用於共享源極區域 14 之兩列記憶體單元的所有源極區域 14 電氣連接在一起。各位元線 16a 將用於彼行記憶體單元的所有汲極區域 16 電氣連接在一起。各抹除閘線 30a 將用於共享抹除閘 30 之兩列記憶體單元的所有抹除閘 30 電氣連接在一起。正如先前架構，個別的記憶體單元可經獨立地程式化及讀取。然而，無法個別地抹除單元。抹除係藉由將一高正

電壓置於抹除閘線 30a 上來執行，其導致同時抹除共享相同抹除閘線 30a 的兩列記憶體單元。例示性操作電壓可包括在下方表 2 中者（在此實施例中，選擇閘線 28a 可稱為字線 WL）：

表 2

	WL		BL		SL		CG		EG	
	選取	非選取	選取	非選取	選取	非選取	選取	非選取	選取	非選取
抹除	0 V	0 V	0 V	0 V	0 V	0 V	0 V	0 V	11.5 V	0 V
讀取	2.5 V	0 V	0.8 V	0 V	0 V	0 V	2.5 V	2.5 V	0 V	0 V
程式化	1 V	0 V	1 μA	2.5 V	4.5 V	0.5 V	10.5 V	0/2.5 V	4.5 V	0.5 V

【0020】 為了利用包含上述類型的非揮發性記憶體單元中的一者之記憶體陣列，進行了兩種修改。第一，該等線經重組態使得各記憶體單元可個別地程式化、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，如下文所進一步解釋者。第二，提供記憶體單元連續（類比）程式化。具體而言，陣列中的各記憶體單元的記憶體狀態（亦即，浮閘上的電荷）可獨立地且對其他記憶體單元的干擾最小地從完全抹除狀態連續地改變至完全程式化狀態，且反之亦然。此意指單元儲存係類比的或至少可儲存許多離散值（諸如 16 或 32 個不同的值）之一者，其允許非常精確及個別的調諧記憶體陣列中的所有單元，且其使記憶體陣列非常適於儲存類神經網路的突觸權重並對該等突觸權重進行微調調整。

記憶體單元程式化及儲存

【0021】 類神經網路權重等級分配在儲存於記憶體單元中時可經均勻地間隔（如圖 8A 所示）或經不均勻地間隔（如圖 8B 所示）。非揮發性記憶體單元的程式化可使用一雙向調諧演算法實施，諸如圖 9 所示者。 I_{cell} 係經程式化之目標單元的讀取電流，且 I_{target} 係當該單元經理想地程式化時所欲的讀取電流。讀取目標單元讀取電流 I_{cell} （步驟 1），並與目標讀取電流 I_{target} 比較（步驟 2）。若目標單元讀取電流 I_{cell} 大於目標讀取電流 I_{target} ，實施程式化調諧程序（步驟 3）以增加浮閘上的電子數目（其中查詢表係用以判定控制閘上所欲的程式化電壓 V_{CG} ）（步驟 3a 至步驟 3b），其可依需要重複（步驟 3c）。若目標單元讀取電流 I_{cell} 小於目標讀取電流 I_{target} ，實施抹除調諧程序（步驟 4）以減少浮閘上的電子數目（其中查詢表係用以判定抹除閘上所欲的抹除電壓 V_{EG} ）（步驟 4a 至步驟 4b），其可根據需要重複（步驟 4c）。若程式化調諧程序過衝(overshoot)目標讀取電流，則實施抹除調諧程序（步驟 3d 並從步驟 4a 開始），且反之亦然（步驟 4d 並從步驟 3a 開始），直到達成目標讀取電流（在可接受的 δ 值內）。

【0022】 非揮發性記憶體單元的程式化可替代地使用其使用程式化調諧的單向調諧演算法實作。以此演算法，記憶體單元最初係完全抹除，然後實施圖 9 中的程式化調諧步驟 3a 至 3c，直到目標單元的讀取電流到達目標臨限值。替代地，非揮發性記憶體單元的調諧可使用其使用抹除調諧的單向調諧演算法實作。在此方法中，記憶體單元

最初係完全程式化，然後執行圖 9 中的抹除調諧步驟 4a 至 4c，直到目標單元的讀取電流到達目標臨限值。

【0023】 圖 10 係繪示使用電流比較之權重映射的圖。權重數位位元（例如，5 位元權重用於各突觸，表示記憶體單元的目標數位權重）係輸入至數位類比轉換器(DAC) 40，其將位元轉換至電壓 V_{out} （例如 64 個電壓位準–5 個位元）。 V_{out} 係藉由電壓至電流轉換器 V/I Conv 42 轉換至電流 I_{out} （例如，64 個電流位準–5 個位元）。電流係供應至電流比較器 IComp 44。程式化或抹除演算法啟用係輸入至記憶體單元 10（例如，抹除：遞增 EG 電壓；或程式化：遞增 CG 電壓）。記憶體單元電流輸出 $I_{cellout}$ （亦即，來自讀取操作）係供應至電流比較器 IComp 44。電流比較器 IComp 44 比較記憶體單元電流 $I_{cellout}$ 與從權重數位位元導出的電流 I_{out} ，以產生指示儲存在記憶體單元 10 中之權重的信號。

【0024】 圖 11 係繪示使用電壓比較之權重映射的圖。權重數位位元（例如，5 位元權重用於各突觸）係輸入至數位類比轉換器(DAC) 40，其將位元轉換至電壓 V_{out} （例如，64 個電壓位準–5 個位元）。 V_{out} 係供應至電壓比較器 VComp 46。程式化或抹除演算法啟用係輸入至記憶體單元 10（例如，抹除：遞增 EG 電壓；或程式化：遞增 CG 電壓）。記憶體單元電流輸出 $I_{cellout}$ 係供應至電流至電壓轉換器 I/V Conv 48，以用於轉換至電壓 V_{2out} （例如，64 個電壓位準–5 個位元）。電壓 V_{2out} 係供應至電壓比較器 VComp 46。電壓比較器

VComp 46 比較電壓 V_{out} 與 $V2_{out}$ 以產生指示儲存在記憶體單元 10 中之權重的信號。

採用非揮發性記憶體單元陣列的類神經網路

【0025】 圖 12 概念性地繪示利用一非揮發性記憶體陣列之一類神經網路的一非限制性實例。此實例將非揮發性記憶體陣列類神經網路使用於面部識別應用，但任何其他適當應用可使用基於非揮發性記憶體陣列的類神經網路實作。S0 係輸入，針對此實例，其係具有 5 位元精確度之 32×32 像素的 RGB 影像（亦即，三個 32×32 像素陣列，一者用於各顏色 R、G、及 B，各像素係 5 位元精確度）。從 S0 到 C1 的突觸 CB1 具有二組不同的權重及共享權重，並以 3×3 像素重疊濾波器（核心），將濾波器移位 1 個像素（或如模型所決定而多於 1 個像素）掃描輸入影像。具體而言，將影像之 3×3 部分中的 9 個像素（亦即，稱為濾波器或核心）的值提供至突觸 CB1，藉此將此等 9 個輸入值乘以合適權重，且在對該乘法的輸出加總之後，由 CB1 的第一神經元判定並提供用於生成特徵圖譜(feature map) C1 之該等層的一者之像素的單一輸出值。然後將 3×3 濾波器右移一個像素（亦即，在右邊加入三個像素之行，並在左邊丟棄三個像素之行），藉此將此新定位濾波器中的 9 個像素值提供至突觸 CB1，藉此將彼等乘以相同權重，並藉由關聯神經元判定第二單一輸出值。繼續此程序直到 3×3 濾波器針對所有三種顏色及所有位元（精確值）掃描整個 32×32 像素影像。然

後使用不同組的權重重複該程序，以生成 C1 之不同的特徵圖譜，直到層 C1 的所有特徵圖譜已計算。

【0026】 在本實例中，在 C1 處，有各具有 30×30 個像素之 16 個特徵圖譜。各像素係從將輸入與核心相乘所擷取的新特徵像素，且因此各特徵圖譜係二維陣列，且因此在此實例中，突觸 CB1 構成二維陣列的 16 層（請記住本文中所提及的神經元層及陣列係邏輯關係，不一定係實體關係—亦即，陣列不一定以實體的二維陣列定向）。16 個特徵圖譜之各者係由經施加至濾波器掃描之十六組不同的突觸權重之一者生成。C1 特徵圖譜可全部關於相同影像特徵的不同態樣，諸如邊界識別。例如，第一圖譜（使用第一組權重生成，用於所有掃描所共享而用以生成此第一圖譜）可識別圓形邊緣，第二圖譜（使用與第一組權重不同的第二組權重生成）可識別矩形邊緣，或某些特徵的縱橫比，並依此類推。

【0027】 啟動功能 P1（彙整）係在從 C1 到 S1 之前施加，其彙整來自各特徵圖譜中之連續、非重疊 2×2 區域的值。彙整階段的目的係平均附近位置（或也能使用 max 函數），以減少，例如，邊緣位置的相依性，並在進入下一階段之前減少資料大小。在 S1 處，有 16 個 15×15 特徵圖譜（亦即，各 15×15 像素的十六個不同陣列）。從 S1 到 C2 之 CB2 中的突觸及關聯神經元使用 4×4 濾波器，以 1 個像素的濾波器移位來掃描 S1 中的圖譜。在 C2 處，有 22 個 12×12 特徵圖譜。啟動功能 P2（彙整）係在從 C2 到 S2 之前施加，其彙整來自各特徵圖譜中之連續、非重疊 2×2 區域的值。在 S2 處，有 22 個 6×6 特徵圖譜。

將啟動功能施加在從 S2 到 C3 的突觸 CB3 處，其中 C3 中的每個神經元連接至 S2 中的每個圖譜。在 C3 處，有 64 個神經元。從 C3 到輸出 S3 的突觸 CB4 將 S3 完全連接至 C3。在 S3 處的輸出包括 10 個神經元，其中最高的輸出神經元判定類別。該輸出可例如指示原始影像之內容的識別或類別。

【0028】 突觸的各級係使用非揮發性記憶體單元的陣列或陣列的一部分實作。圖 13 係包括非揮發性記憶體單元且係用作一輸入層與次一層之間的突觸之向量矩陣乘法(vector-by-matrix multiplication, VMM)陣列的方塊圖。具體而言，VMM 32 包括非揮發性記憶體 33 的陣列、抹除閘及字線閘解碼器 34、控制閘解碼器 35、位元線解碼器 36、及源極線解碼器 37，其等解碼記憶體陣列 33 的輸入。此實例中的源極線解碼器 37 也解碼記憶體單元陣列的輸出。記憶體陣列用作兩個目的。第一，其儲存將由 VMM 使用的權重。第二，記憶體陣列有效地將輸入乘以儲存在記憶體陣列中的權重以產生輸出，其將係至次一層的輸入或至最終層的輸入。藉由執行乘法功能，記憶體陣列否定對分開的乘法邏輯電路的需求，且亦由於原位記憶體運算而係有能源效率的。

【0029】 將記憶體陣列的輸出供應至一差分加算器（諸如加總運算放大器）38，其加總記憶體單元陣列的輸出以創建用於該卷積的單一值。該差分加算器係諸如用以實現正權重及負權重與正輸入的加總。然後將該經加總輸出值供應至整流輸出的啟動功能電路 39。啟動功能可包括 S 型(sigmoid)函數、雙曲正切(tanh)函數、ReLU 函數。經整流

輸出值變成作為次一層（例如，以上描述中的 C1）之特徵圖譜的元素，然後經施加於次一突觸以產生次一特徵圖譜層或最終層。因此，在此實例中，記憶體陣列構成複數個突觸（其從神經元的先前層或從輸入層（諸如影像資料庫）接收其等的輸入），且加總運算放大器 38 及啟動功能電路 39 構成複數個神經元。

【0030】 圖 14 係 VMM 之各種級的方塊圖。如圖 14 所示，輸入係由數位類比轉換器 31 來從數位轉換成類比，並提供至輸入 VMM 32a。將由輸入 VMM 32a 生成的輸出提供作為至次一 VMM（隱藏級 1）32b 的輸入，其繼而生成提供作為至次一 VMM（隱藏級 2）32b 之輸入的輸出，並依此類推。VMM 32 之各種層如卷積式類神經網路 (CNN) 之突觸及神經元的不同層而作用。各 VMM 可係獨立的非揮發性記憶體陣列，或多個 VMM 可利用相同的非揮發性記憶體陣列的不同部分，或多個 VMM 可利用相同的非揮發性記憶體陣列的重疊部分。

【0031】 圖 15 繪示四閘極記憶體單元（亦即，諸如圖 6 所示者）的一陣列，其經配置為一汲極加總矩陣乘法器。用於圖 15 之陣列的各種閘極及區域線與圖 7 中者相同（對於對應結構具有相同的元件編號），除了抹除閘線 30a 垂直地而非水平地延行（亦即，各抹除閘線 30a 將用於記憶體單元之該行的所有抹除閘 30 連接在一起）以外，使得各記憶體單元 10 可獨立地經程式化、抹除、及讀取。在記憶體單元之各者以該單元的適當權重值程式化後，該陣列充當汲極加總矩陣乘法器。矩陣輸入係 $V_{in0} \dots V_{in7}$ ，並係置放於選擇閘線 28a 上。用於圖 15 的陣列之輸出的矩陣 $I_{out0} \dots I_{outN}$ 係產生在位元線 16a 上。對於該

行中的所有單元，各輸出 I_{out} 係單元電流 I 乘以儲存在該單元中之權重 W 的和：

$$I_{out} = \sum (I_{ij} * W_{ij})$$

【0032】 各記憶體單元（或記憶體單元對）充當具有表示為輸出電流 I_{out} 之權重值的單一突觸，該輸出電流藉由儲存在該行中之記憶體單元（或記憶體單元對）中的權重值之和決定。任何給定突觸的輸出係以電流的形式。因此，第一階段之後的各後續 **VMM** 階段較佳地包括用於將來自先前 **VMM** 階段的進入電流轉換成待使用作為輸入電壓 V_{in} 之電壓的電路系統。圖 16 繪示此類電流至電壓轉換電路系統的一實例，該電流至電壓轉換電路系統係將進入電流 $I_{in0} \cdots I_{inN}$ 對數轉換為輸入電壓 $V_{in0}..V_{inN}$ 之記憶體單元的一經修改列。

【0033】 本文中描述的記憶體單元以弱反轉偏壓，

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

$$w = e^{(-V_{th})/kV_t}$$

對於使用記憶體單元以將輸入電流轉換成輸入電壓的 **I** 至 **V** 對數轉換器：

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

對於使用作為向量矩陣乘法器 **VMM** 的記憶體陣列，輸出電流係：

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

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

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

【0034】 圖 17 及圖 18 繪示四閘極記憶體單元（亦即，諸如圖 6 所示者）之一陣列的另一組態，其經配置為一汲極加總矩陣乘法器。圖 17 及圖 18 之陣列的線與圖 15 及圖 16 之陣列中者相同，除了源極線 14a 垂直地而非水平地延行（亦即，各源極線 14a 將記憶體單元之該行的所有源極區域 14 連接在一起），及抹除閘線 30a 水平地而非垂直地延行（亦即，各抹除閘線 30a 將記憶體單元對之該列的所有抹除閘 30 連接在一起）以外，使得各記憶體單元可獨立地經程式化、抹除、及讀取。矩陣輸入 $V_{in0}...V_{inN}$ 保留在選擇閘線 28a 上，且矩陣輸出 $I_{out0}...I_{outN}$ 保留在位元線 16a 上。

【0035】 圖 19 繪示四閘極記憶體單元（亦即，諸如圖 6 所示者）之一陣列的另一組態，其經配置為一閘極耦接/源極加總矩陣乘法器。用於圖 19 之陣列的線與圖 15 及圖 16 中者相同，除了選擇閘線 28a 垂直地延行且記憶體單元的各行有其等的二者。具體而言，記憶體單元的各行包括兩個選擇閘線：第一選擇閘線 28a1 將奇數列記憶體單元的所有選擇閘 28 連接在一起，且第二選擇閘線 28a2 將偶數列記憶體單元的所有選擇閘 28 連接在一起。

【0036】 在圖 19 之頂部及底部的電路用以將輸入電流 $I_{in0}...I_{inN}$ 對數轉換成輸入電壓 $V_{in0}..V_{inN}$ 。顯示於此圖中的矩陣輸入係 $V_{in0}...V_{in5}$ 並經置放於選擇閘線 28a1 及 28a2 上。具體而言，將輸入 V_{in0} 經置放於行 1 中之奇數單元的選擇線 28a1 上。 V_{in1} 係置放

於行 1 中之偶數單元的選擇閘線 28a2 上。 V_{in2} 係置放於行 2 中之奇數單元的選擇閘線 28a1 上。 V_{in3} 係置放於行 2 中之偶數單元的選擇閘線 28a2 上，並依此類推。將矩陣輸出 $I_{out0}...I_{out3}$ 提供在源極線 14a 上。位元線 16a 以固定偏壓電壓 V_{BLrd} 而偏壓。對於記憶體單元之該列中的所有單元，各輸出 I_{out} 係單元電流 I 乘以儲存在該單元中之權重 W 的和。因此，對於此架構，記憶體單元的各列充當具有表示為輸出電流 I_{out} 之權重值的單一突觸，該輸出電流藉由儲存在該列中之記憶體單元中的權重值之和決定。

【0037】圖 20 繪示四閘極記憶體單元（亦即，諸如圖 6 所示者）之一陣列的另一組態，其經配置為一閘極耦接/源極加總矩陣乘法器。用於圖 20 之陣列的線與圖 19 中者相同，除了位元線 16 垂直地延行且記憶體單元的各行有其等的二者。具體而言，記憶體單元的各行包括兩個位元線：將相鄰的雙記憶體單元（共享相同的位元線接觸件的二個記憶體單元）的所有汲極區域連接在一起之第一位元線 16a1，及將次一相鄰的雙記憶體單元的所有汲極區域連接在一起之第二位元線 16a2。矩陣輸入 $V_{in0}...V_{inN}$ 保留在選擇閘線 28a1 及 28a2 上，且矩陣輸出 $I_{out0}...I_{outN}$ 保留在源極線 14a 上。所有第一位元線 16a1 的組係以一偏壓位準（例如，1.2 v）來偏壓，且所有第二位元線 16a2 的組係以另一偏壓位準（例如，0 v）來偏壓。源極線 14a 以虛擬偏壓位準（例如，0.6 v）來偏壓。對於共享共同源極線 14a 的各對記憶體單元，輸出電流將係頂部單元減底部單元的差動輸出。因此，各輸出 I_{out} 係此等差動輸出的和：

$$I_{out} = \Sigma (I_{iju} * W_{iju} - I_{ijd} * W_{ijd})$$

$$SL \text{ 電壓} \sim \frac{1}{2} V_{dd}, \sim 0.6 \text{ v}$$

因此，對於此架構，成對記憶體單元的各列充當具有表示為輸出電流 I_{out} 之權重值的單一突觸，該輸出電流係由儲存在成對記憶體單元的該列中之記憶體單元中的權重值（例如，一個正權重及一個負權重）所決定之差動輸出的和。

【0038】圖 21 繪示四閘極記憶體單元（亦即，諸如圖 6 所示者）之一陣列的另一組態，其經配置為一閘極耦接/源極加總矩陣乘法器。用於圖 21 之陣列的線與圖 20 中者相同，除了抹除閘 30a 水平地延行，且控制閘線 22a 垂直地延行且記憶體單元的各行有其等的二者。具體而言，記憶體單元的各行包括兩個控制閘線：第一控制閘線 22a1 將奇數列記憶體單元的所有控制閘 22a 連接在一起，且第二控制閘線 22a2 將偶數列記憶體單元的所有控制閘 22a 連接在一起。矩陣輸入 $V_{in0} \dots V_{inN}$ 保留在選擇閘線 28a1 及 28a2 上，且矩陣輸出 $I_{out0} \dots I_{outN}$ 保留在源極線 14a 上。

【0039】圖 22 繪示四閘極記憶體單元（亦即，諸如圖 6 所示者）之一陣列的另一組態，其經配置為一源極加總矩陣乘法器。用於圖 22 之陣列的線及輸入與圖 17 中者相同。然而，將其等提供在源極線 14a 上，而不是將輸出提供在位元線 16a 上。矩陣輸入 $V_{in0} \dots V_{inN}$ 保留在選擇閘線 28a 上。

【0040】圖 23 繪示二閘極記憶體單元（亦即，諸如圖 1 所示者）之一陣列的一組態，其經配置為一汲極加總矩陣乘法器。用於圖 23 之陣列的線與圖 5 中者相同，除了水平源極線 14a 已以垂直源極線 14a 置換。具體而言，各源極線 14a 經連接至記憶體單元之該行中的所有源極區域。矩陣輸入 $V_{in0}...V_{inN}$ 係置放於控制閘線 22a 上。矩陣輸出 $I_{out0}...I_{outN}$ 係產生在位元線 16a 上。對於該行中的所有單元，各輸出 I_{out} 係單元電流 I 乘以儲存在該單元中之權重 W 的和。記憶體單元的各行充當具有表示為輸出電流 I_{out} 之權重值的單一突觸，該輸出電流藉由儲存在該行中之記憶體單元中的權重值之和決定。

【0041】圖 24 繪示二閘極記憶體單元（亦即，諸如圖 1 所示者）之一陣列的一組態，其經配置為一源極加總矩陣乘法器。用於圖 24 之陣列的線與圖 5 中者相同，除了控制閘線 22a 垂直地延行且記憶體單元的各行有其等的二者。具體而言，記憶體單元的各行包括兩個控制閘線：第一控制閘線 22a1 將奇數列記憶體單元的所有控制閘 22a 連接在一起，且第二控制閘線 22a2 將偶數列記憶體單元的所有控制閘 22a 連接在一起。

【0042】此組態的矩陣輸入係 $V_{in0}...V_{inN}$ 並經置放於控制閘線 22a1 及 22a2 上。具體而言，輸入 V_{in0} 係置放於行 1 中之奇數列單元的控制閘線 22a1 上。 V_{in1} 係置放於行 1 中之偶數列單元的控制閘線 22a2 上。 V_{in2} 係置放於行 2 中之奇數列單元的控制閘線 22a1 上。 V_{in3} 係置放於行 2 中之偶數列單元的控制閘線 22a2 上，並依此類推。矩陣輸出 $I_{out0}...I_{outN}$ 係產生在源極線 14a 上。對於共享共同源極線

14a 的各對記憶體單元，輸出電流將係頂部單元減底部單元的差動輸出。因此，對於此架構，成對記憶體單元的各列充當具有表示為輸出電流 I_{out} 之權重值的單一突觸，該輸出電流係由儲存在成對記憶體單元的該列中之記憶體單元中的權重值所決定之差動輸出的和。

【0043】 圖 15 至圖 16、圖 19、及圖 20 之實施例的例示性操作電壓包括：

	EG		WL		CG		BL		SL	
	選取	非選取	選取	非選取	選取	非選取	選取	非選取	選取	非選取
抹除	VEGerase	gnd	gnd	gnd	gnd	VCGerinhbit	gnd	gnd	gnd	gnd
程式化	VEGprg/ gnd	gnd	VWLprg	gnd	VCGprg	gnd	Iprog	VBLprginh	VSprg	gnd
讀取	gnd	gnd	VWLrd	gnd	VCGrd	gnd	VBLrd	gnd	VSrd	float/ gnd

近似數值包括：

VEGerase	8 至 11.5 v
VCGerinhbit	3.5 至 8 v
VEGprg	4 至 6 v
VWLprg	0.8 至 1.2 v
VCGprg	6 至 10 v
VBLprginh	1 至 2.5 v
Iprog	0.2 至 1 μ a
VSprg	3 至 5 V
VWLrd	0.4 至 2.0 V
VCGrd	0 至 2.5 V
VBLrd	1 至 2 V
VSrd	0 至 0.6 V

【0044】圖 17 至圖 18 及圖 22 之實施例的例示性操作電壓包括：

	EG		WL		CG			BL		SL	
	選取	非選取	選取	非選取	選取	非選取- 共享 EG	非選取	選取	非選取	選取	非選取
抹除	VEGerase	gnd	gnd	gnd	gnd	VCGerinh	gnd	gnd	VBLerinh/ gnd	VSlerinh	gnd
程式化	VEGprg/ gnd	gnd	VWLprg	gnd	VCGprg/ gnd	gnd	gnd	Iprog	VBLprginh	VSprg	gnd
讀取	gnd	gnd	VWLrd	gnd	VCGrd	gnd	gnd	VBLrd	gnd	VSrd	浮接/gnd

近似數值包括：

VEGerase	7 至 10 v
VSlerinh	3.5 至 6 v
VCGerinh	3.5 至 7 v
VBLerinh	1 至 2.5 v
VEGprg	4 至 6 v
VWLprg	0.8 至 1.2 V
VCGprg	6 至 10 v
VBLprginh	1 至 2.5 v
Iprog	0.2 至 1 μa
VSprg	3 至 5 V
VWLrd	0.4 至 2.0 V
VCGrd	1 至 2.5 V
VBLrd	1 至 2 V
VSrd	0 至 0.6 V

【0045】圖 25 繪示用於與本發明併用之一例示性電流至電壓對數轉換器 50（WL=選擇閘線、CG=控制閘線、EG=抹除閘線）。記憶體係偏壓在弱反轉區中， $I_{ds} = I_o * e^{(V_g - V_{th})/kV_t}$ 。圖 26 繪示用於與本發明併用之一例示性電壓至電流對數轉換器 52。記憶體係偏壓在弱反轉區中。圖 27 繪示用於與本發明併用之一參考 Gnd 的電流加算器 54。

圖 28 以下繪示用於與本發明併用之一參考 Vdd 的電流加算器 56。負載的實例包括二極體、非揮發性記憶體單元、及電阻器。

【0046】 上述記憶體陣列組態實作前饋的分類引擎。訓練係藉由將「權重」值儲存在記憶體單元（創建突觸陣列）中而完成，其意指個別單元的次臨限斜率因子已被修改。神經元係藉由加總突觸之輸出及取決於神經元臨限（例如，做決策）排除或不排除而實作。

【0047】 以下步驟可用以處理輸入電流 I_E （例如，輸入電流直接來自用於影像識別之特徵計算的輸出）：

步驟 1-轉換為更容易以非揮發性記憶體處理的對數尺度。

- 使用雙極電晶體的輸入電流至電壓轉換。雙極電晶體的偏壓電壓 V_{BE} 與射極電流有對數關係。
- $V_{BE} = a \cdot \ln I_E - b \rightarrow V_{BE} \propto \ln I_E$
- 其中 a （比率）及 b （偏壓或偏移）係常數
- 生成 V_{BE} 電壓使得記憶體單元將在次臨限區域中操作。

步驟 2-將所生成的偏壓 V_{BE} 施加至字線（在次臨限區域中）。

- CMOS 電晶體的輸出電流 I_{DRAIN} 與輸入電壓(V_{GS})、熱電壓(U_T)、及 κ ($\kappa = C_{ox}/(C_{ox}+C_{dep})$)有指數關係，其中 C_{ox} 及 C_{dep} 線性取決於浮閘上的電荷。
- $I_{DRAIN} \propto \text{Exp}(kV_{BE} / U_T)$ ，或
- $\ln I_{DRAIN} \propto kV_{BE}/U_T$
- I_{DRAIN} 的對數與 V_{BE} 的倍數及浮閘上的電荷（相關於 κ ）有線性關係，其中 U_T 在指定溫度係常數。

- 輸出=輸入*權重的關係對於突觸存在。

【0048】 單元之各者的輸出(I_{DRAIN})在讀取模式中可連結在一起以加總陣列或陣列之扇區中之各突觸的值。一旦 I_{DRAIN} 已經加總，其可經饋送至電流比較器中，並取決於單一感知類神經網路的比較而輸出「邏輯」0 或 1。一感知（一扇區）已於上文描述。來自各感知的輸出可經饋送到用於多個感知的次一組扇區。

【0049】 在基於記憶體體的卷積式類神經網路中，一組輸入需要乘以某些權重以產生隱藏層或輸出層的所欲結果。如上文所解釋的，一種技術係使用 $M \times M$ 濾波器（核心）（例如， $N \times N$ 矩陣掃描先前影像，該 $M \times M$ 濾波器在水平及垂直方向兩者中跨該影像移位 X 個像素。只要有足夠的輸入至記憶體陣列，像素的掃描可至少部分地並行完成。例如，如圖 29 所示，可使用 $M=6$ 的濾波器尺寸（亦即，36 個像素的 6×6 陣列）以使用 $X=2$ 的移位掃描一 $N \times N$ 影像陣列。在該實例中，將濾波器中之第一列的六個像素提供至對 N^2 個輸入之記憶體陣列的前 6 個輸入。然後，將濾波器中之第二列的六個像素提供至 N^2 個輸入之第二 N 個輸入中的前 6 個輸入，並依此類推。此係表示在圖 29 中之圖的第一列中，其中該等點表示如上文所陳述之用於與輸入的乘法之儲存在記憶體陣列中的權重。然後，濾波器向右移位二個像素，並將已移位濾波器中之第一列的六個像素提供至第一 N 個輸入的第三至第八輸入，將第二列的六個像素提供至第二 N 個輸入的第三至第八輸入，並依此類推。一旦將濾波器完全移位到影像右側，濾波器係重定位回到左側，但向下移位二個像素，其中該程序再次重複，直到掃描整個

$N \times N$ 影像。各組水平移位掃描可以梯形形狀表示，該梯形形狀顯示 N^2 個記憶體陣列輸入的何者具備用於乘法的資料。

【0050】據此，使用掃描間的二個像素移位及 6×6 的濾波器尺寸， $N \times N$ 影像陣列的掃描需要 N^2 個輸入及 $((N-4)/2))^2$ 列。圖 30 圖形地顯示梯形形狀，其指示記憶體陣列中的權重如何針對濾波器掃描儲存。各列的陰影區代表在一組水平掃描期間施加於輸入的權重。箭頭指示記憶體陣列的線性輸入線（例如，在圖 15 中之接收輸入資料的輸入線 28a 以線性方式完全延伸跨記憶體陣列，各一者總是存取記憶體單元的相同列；在圖 19 之陣列的情況中，輸入線之各者總是存取記憶體單元的相同行）。白色區指示無資料提供給輸入之處。因此，白色區域指示記憶體單元陣列的無效率使用。

【0051】藉由如圖 31 所示地重組態記憶體陣列，可增加效率，並減少輸入的總數。具體而言，將記憶體陣列的輸入線週期地移位至另一列或行，因此減少陣列的未使用部分，並因此減少實施掃描所需之陣列上的重複輸入線的數目。具體而言，在本實例之移位 $X=2$ 的情況中，箭頭指示各輸入線週期地移位兩列或兩行，將寬間隔記憶體單元利用梯形形狀轉換成緊密間隔記憶體單元利用矩形形狀。雖然在記憶體單元部分之間需要用於導線束的額外空間以實作此移位，但記憶體單元陣列中所需要的輸入數目大幅減少（僅 $5n+6$ ）。

【0052】圖 32 繪示圖 15 的陣列，但對於用作輸入線的線 28a 具有兩列的週期性移位。輸入線之列中的週期性移位可類似地實作在圖 17、圖 22、及圖 23 的陣列中。圖 33 繪示圖 20 的陣列，但對於用作

輸入線的線 28a1 及 28a2 具有兩行的週期性移位。輸入線之行中的週期性移位可類似地實作在圖 19、圖 21、及圖 24 的陣列中。

改善的調諧機制及演算法

【0053】 現將描述用於改善的調諧機制及演算法的實施例。調諧係如下的程序：藉其驗證所欲的電荷量經儲存在一非揮發性記憶體單元的浮閘中，也就是，確保非揮發性記憶體單元儲存所欲的值。

【0054】 圖 35 描繪神經元 VMM 3500，其特別適於圖 2 所示之類型的記憶體單元，且係用作一輸入層與次一層之間的突觸。VMM 3500 包含非揮發性記憶體單元的記憶體陣列 3503、參考陣列 3501、及參考陣列 3502。參考陣列 3501 及 3502 用以將流入端子 BLR0 至 3 中的電流輸入轉換成電壓輸入 WL0 至 3。實際上，參考記憶體單元係透過多工器與流入其等的電流輸入連接的二極體。參考單元經調諧至目標參考位準。記憶體陣列 3503 供兩個目的所用。第一，其儲存將由 VMM 3500 使用的權重。第二，記憶體陣列 3503 有效地將輸入（端子 BLR0 至 3 中所提供的電流輸入；參考陣列 3501 及 3502 將此等電流輸入轉換成輸入電壓以供應給字線 WL0 至 3）乘以儲存在記憶體陣列中的權重以產生輸出，該輸出將係至次一層的輸入或至最終層的輸入。藉由實施乘法功能，記憶體陣列否定對分開的乘法邏輯電路的需求，且也係有能源效率的。此處，電壓輸入係在字線上提供，且輸出在一讀取（推論）操作期間顯現於位元線上。置於位元線上的電流執行來自連接至位元線之記憶體單元的所有電流之一加總功能。

【0055】 圖 36 描繪用於 VMM 3500 的操作電壓。表中的行指示置於用於經選擇單元之字線、用於未經選擇單元之字線、用於經選擇單元之位元線、用於未經選擇單元之位元線、用於經選擇單元之源極線、以及用於未經選擇單元之源極線上的電壓。列指示讀取、抹除、及程式化操作。

【0056】 圖 37 描繪神經元 VMM 3700，其特別適於圖 2 所示之類型的記憶體單元，且係用作一輸入層與次一層之間的突觸。VMM 3700 包含非揮發性記憶體單元的記憶體陣列 3703、參考陣列 3701、及參考陣列 3702。VMM 3700 類似於 VMM 3500，除了在 VMM 3700 中，字線在垂直方向上延行。此處，輸入係在字線上提供，且輸出在一讀取操作期間顯現於源極線上。置於源極線上的電流執行來自連接至源極線之記憶體單元的所有電流之一加總功能。

【0057】 圖 38 描繪用於 VMM 3700 的操作電壓。表中的行指示置於用於經選擇單元之字線、用於未經選擇單元之字線、用於經選擇單元之位元線、用於未經選擇單元之位元線、用於經選擇單元之源極線、以及用於未經選擇單元之源極線上的電壓。列指示讀取、抹除、及程式化操作。

【0058】 圖 39 描繪神經元 VMM 3900，其特別適於圖 6 所示之類型的記憶體單元，且係用作一輸入層與次一層之間的突觸。VMM 3900 包含非揮發性記憶體單元的記憶體陣列 3903、參考陣列 3901、及參考陣列 3902。參考陣列 3901 及 3902 用以將流入端子 BLR0 至 3 中的電流輸入轉換成電壓輸入 CG0 至 3。實際上，參考記憶體單元係

透過多工器與流入其等的電流輸入經二極體連接。參考單元經調諧至目標參考位準。記憶體陣列 3903 供兩個目的所用。第一，其儲存將由 VMM 3900 使用的權重。第二，記憶體陣列 3903 有效地將輸入（提供給端子 BLR0 至 3 的電流輸入；參考陣列 3901 及 3902 將此等電流輸入轉換成輸入電壓以供應給控制閘 CG0 至 3）乘以儲存在記憶體陣列中的權重以產生輸出，該輸出將係至次一層的輸入或至最終層的輸入。藉由實施乘法功能，記憶體陣列否定對分開的乘法邏輯電路的需求，且也係有能源效率的。此處，輸入係在字線上提供，且輸出在一讀取操作期間顯現於位元線上。置於位元線上的電流執行來自連接至位元線之記憶體單元的所有電流之一加總功能。

【0059】 VMM 3900 針對記憶體陣列 3903 中的記憶體單元實施單向調諧。也就是，各單元經抹除，並接著經部分程式化，直到達到浮閘上的所欲電荷。若將太多電荷置於浮閘上（使得錯誤的值經儲存在單元中），則必須抹除單元，且部分程式化操作的序列必須重新開始。如所示，共享相同抹除閘的兩列需要一起抹除（已知為頁面抹除），其後，各單元經部分程式化直到達到浮閘上的所欲電荷，

【0060】 圖 40 描繪用於 VMM 3900 的操作電壓。表中的行指示置於用於經選擇單元之字線、用於未經選擇單元之字線、用於經選擇單元之位元線、用於未經選擇單元之位元線、用於經選擇單元之控制閘、用於與經選擇單元相同扇區中之未經選擇單元之控制閘、用於不同於經選擇單元之一扇區中之未經選擇單元之控制閘、用於經選擇單元之抹除閘、用於未經選擇單元之抹除閘、用於經選擇單元之源極線、

用於未經選擇單元之源極線上的電壓。列指示讀取、抹除、及程式化操作。

【0061】 圖 41 描繪神經元 VMM 4100，其特別適於圖 6 所示之類型的記憶體單元，且係用作一輸入層與次一層之間的突觸。VMM 4100 包含非揮發性記憶體單元的記憶體陣列 4103、參考陣列 4101、及參考陣列 4102。除了 VMM 4100 實施雙向調諧以外，VMM 4100 類似於 VMM 3900，其中各個別單元可經完全抹除、部分程式化、及依需要部分抹除以達到浮閘上所欲的電荷量。如所示，參考陣列 4101 及 4102 將端子 BLR0 至 3 中的輸入電流轉換成欲在列方向上施加至記憶體單元之控制閘電壓 CG0 至 3（通過透過多工器經二極體式連接之參考單元的動作）。電流輸出（神經元）係在位元線中，其將來自連接至位元線之記憶體單元的所有電流加總。

【0062】 圖 42 描繪用於 VMM 4100 的操作電壓。表中的行指示置於用於經選擇單元之字線、用於未經選擇單元之字線、用於經選擇單元之位元線、用於未經選擇單元之位元線、用於經選擇單元之控制閘、用於與經選擇單元相同扇區中之未經選擇單元之控制閘、用於不同於經選擇單元之一扇區中之未經選擇單元之控制閘、用於經選擇單元之抹除閘、用於未經選擇單元之抹除閘、用於經選擇單元之源極線、用於未經選擇單元之源極線上的電壓。列指示讀取、抹除、及程式化操作。

【0063】 圖 43 描繪調諧演算法 4300。調諧演算法 4300 辨識個別的記憶體單元可係快速或慢速。一快速單元係可迅速程式化的一者，

而一慢速單元或一正規單元係相較於快速單元在一程式化操作期間耗費較大量時間來達到相同狀態的一者。此差別係歸因於個別單元特性及差異的物理。

【0064】 調諧演算法 4300 包含下列步驟序列。第一，一記憶體單元頁面經抹除（步驟 4301）。系統接著基於查找表 4320 判定記憶體陣列中的哪些單元係快速單元，該查找表係在一組態序列期間聚集且若用於一單元的特性改變則可在操作期間更新（從快速至慢速或慢速至快速）（步驟 4302）。查找表 4320 可包括例如所有快速非揮發性記憶體單元的一位址列表。或者其可含有用於陣列中之各單元的一項目，且該項目可係單一位元，其中「1」指示一快速單元而「0」指示一慢速正規單元。

【0065】 若一單元係一快速單元，則實施一快速調諧演算法，其中一相對大電荷係透過一部份程式化操作來添加至快速單元的浮閘（步驟 4303）。在各部份程式化操作之後，執行一驗證序列以判定在一讀取操作中通過單元的 I_{cell} 是否大於 $I_{target\ 1}$ （步驟 4304）。若否，則再次執行部份程式化操作。若是，則判定是否 $I_{cell} < I_{margin_0V}$ （步驟 4305）。若是，則在記憶體單元中已達到所欲狀態，且調諧序列完成（步驟 4306）。若否，則單元已較預期更快地程式化，且其係在查找表 4320 中標示為一快速單元（步驟 4307）。由於太多電荷已置於浮閘上，單元不使用，且其必須再次經抹除（步驟 4302）。

【0066】 若步驟 4302 斷定一單元係一慢速單元，則實施一低調諧演算法，其中一較小電荷係透過一部份程式化操作來添加至慢速單

元的浮閘（步驟 4308）。在各部分程式化操作之後，執行一驗證序列以判定在一讀取操作中通過單元的 I_{cell} 是否大於 $I_{target\ 1}$ （步驟 4309）。若否，則再次執行部分程式化操作。若是，則判定是否 $I_{cell} < I_{margin_0V}$ （步驟 4310）。若是，則在記憶體單元中已達到所欲狀態，且調諧序列完成（步驟 4311）。若否，則單元已較預期更快地程式化，且其係在查找表 4320 中標示為一快速單元（步驟 4307）。由於太多電荷已置於浮閘上，單元不使用，且其必須再次經抹除（步驟 4302）。快速調諧演算法可使用一大寫入（例如程式化）電壓增量或寬寫入脈衝寬度來實施，低調諧演算法可使用一小寫入電壓增量或一窄寫入脈衝寬度來實施。

【0067】 圖 44 描繪調諧演算法 4400。調諧演算法 4400 可在一組態序列期間使用以識別一記憶體陣列中係快速單元之單元。調諧演算法 4400 包含下列步驟序列。一頁面經抹除（步驟 4401）。以一電壓 $V_{CG_diagnosis}$ 程式化一單元（步驟 4402）。電流 I_{cell1} 係在一讀取操作期間通過該單元測量（步驟 4403）。以一電壓 $V_{CG_diagnosis} + dV$ 程式化該單元（步驟 4404）。電流 I_{cell2} 係在一讀取操作期間通過該單元測量（步驟 4405）。就 I_{cell2} 與 I_{cell1} 之間的差值是否超過 I_{cell_0V} 作出一判定（步驟 4406）。若否，則該單元係一正規或慢速單元（步驟 4407）。若是，則該單元係一快速單元，且其係在查找表 4320 中識別為一快速單元（步驟 4408）。如圖 44 中所含的曲線圖所示，一快速單元將回應於步驟 4404 而在其浮閘上累積電荷。增量程式化電壓 d 係一小電壓，且將不影響針對正規或慢速單元的累積電荷。

也就是，快速單元回應於相對小的程式化電壓而在其等之浮閘上累積電荷。替代地，可使用兩個抹除脈衝以提取落入快速位元區域中之記憶體單元的差異電流。在此情況下，首先對所有單元進行深度程式化。

【0068】 圖 45 描繪調諧演算法 4500。調諧演算法 4500 結合粗略演算法 4510（其中一單元係以相對大增量程式化）及精細演算法 4511（其中該單元係以相對小增量程式化）。此最小化過衝所欲電壓的發生並增加系統的總體速度。

【0069】 調諧演算法 4500 包含下列步驟序列。一頁面經抹除（步驟 4501）。接著執行粗略演算法 4510，其包含步驟 4502 及 4503。以 $VCG-C_{init} + dV-C$ 、 $N_{pulseC}=NCi+1$ 程式化一單元（步驟 4502）。電流 I_{cell} 係在一讀取操作期間通過該單元測量，並就是否 $I_{cell} > I_{cell_offset}$ 作出一判定（步驟 4503）。若是，則精細演算法 4511 開始，其包含步驟 4504 及 4505。以 $VCGF_{init} (=VCG-C_{last} - V_{step}) + dV-F$ 、 $N_{pulseF}=NFi+1$ 程式化該單元（步驟 4504）。電流 I_{cell} 係在一讀取操作期間通過該單元測量，並就是否 $I_{cell} > I_{cell_target}$ 作出一判定（步驟 4505）。若是，則已達到所欲電荷，且調諧程序完成（步驟 4506）。若否，則判定是否 $f N_{pulseF} = N_{max}$ （步驟 4507）。若是，則該單元經判定係一不良單元並在一查找表中如此標示（步驟 4508）。若否，則重複步驟 4504。若步驟 4503 的結果為否，則判定是否 $N_{pulseC} = NC_{max}$ （步驟 4509）。若是，則該單元經判定係一不良單元並在一查找表中如此標示（步驟 4508）。若

否，則重複步驟 4502。替代地，可使用寬寫入脈衝寬度代替粗略電壓位準，並可使用窄寫入脈衝寬度代替精細電壓位準。

【0070】 現將參照圖 46 至 50 描述關於圖 43 及圖 45 之調諧操作的額外細節。

【0071】 圖 46 描繪用於均勻步階演算法 4600 的例示性值。在此實例中，近似的程式化目標電壓係 8 V。此目標電壓係例如提取自一查找表或電流目標 I_{target} 對目標程式化電壓的一模型。初始以 4 V 程式化該單元。其後，精細程式化包含使用最大為 400 的步階數以 0.01 V 的步階大小進行程式化。

【0072】 圖 47 描繪用於均勻對數步階（除以十）、粗略/精細演算法 4700 的例示性值。在此實例中，近似的程式化目標電壓係 8 V。初始以 4 V 程式化該單元。其後，粗略程式化包含使用最大為 10 的步階數以 0.4 V ($=4 \text{ V}/10$) 的對數步階大小進行程式化。其後，發生一第一精細程式化序列，其中一部分抹除操作使浮閘電壓減少 0.2 V，且其後，程式化使用最大為 6 的步階數以 0.1 V 的步階大小發生。接著，發生一第二精細程式化序列，其中發生 0.1 V 的一部分抹除操作，且其後，程式化使用最大為 20 的步階數以 0.01 V ($=0.1 \text{ V}/10$) 的對數步階大小發生。因此，脈衝總數係 46。

【0073】 圖 48 描繪二元搜尋步階、粗略/精細演算法 4800。在此實例中，近似的程式化目標電壓再次係 8 V。初始以 4 V 程式化該單元。其後，粗略程式化以差異程式化電壓除以二（即 2 V、1 V、0.5 V、0.25 V、0.125 V、及 0.625 V）的步階發生。其後，執行一抹

除步驟使浮閘電壓減少 0.0625 V，且其後，精細程式化使用最大為 19 的步階數以 0.01 V 的一固定增量發生。此脈衝總數係 25。

【0074】 圖 49 描繪演算法 4800 的一例示性波形。如可見到的，粗略程式化使電壓增加，直到 Icell 超過一臨限。接著執行降壓使電壓減少 0.0625 V，且其後發生精細程式化。降壓係欲避免次一程式化脈衝之潛在的程式化過衝。

【0075】 圖 50 描繪用於本文所述之粗略-精細演算法之另一例示性波形。首先，發生粗略程式化。第二，發生精細程式化。第三，一旦達到適當電壓，記錄程式化電荷的總量及/或記錄步階，使得其後，可以所欲的精確電荷使用單一恆定電壓程式化操作達到適當的程式化位準。替代地，調諧演算法可由下列組成：由寬脈衝寬度及/或大電壓增量組成的粗略脈衝以及由恆定脈衝寬度及/或恆定電壓脈衝組成的精細脈衝。

【0076】 在替代實施例中，圖 15、圖 16、圖 17、圖 18、圖 19、圖 20、圖 21、圖 22、圖 25、圖 26、圖 32、及圖 33 的實施例可經修改使得輸入係控制閘而非選擇閘（字線）。類似地，圖 39 及圖 41 的實施例可經修改使得輸入係選擇閘（字線）而非控制閘。

【0077】 應注意的是，如本文中所使用，「在…上方(over)」及「在…上(on)」之用語皆含括性地包括了「直接在…之上(directly on)」（無居中的材料、元件或間隔設置於其間）及「間接在…之上(indirectly on)」（有居中的材料、元件或間隔設置於其間）的含意。同樣地，「相鄰的(adjacent)」一詞包括了「直接相鄰的」（無居中的

材料、元件或間隔設置於其間)及「間接相鄰的」(有居中的材料、元件或間隔設置於其間)的含意,「安裝於(mounted to)」一詞則包括了「直接安裝於(directly mounted to)」(無居中的材料、元件或間隔設置於其間)及「間接安裝於(indirectly mounted to)」(有居中的材料、元件或間隔設置於其間)的含意,以及「電耦接(electrically coupled)」一詞則包括了「直接電耦接(directly electrically coupled to)」(無居中的材料或元件於其間將各元件電性相連接)及「間接電耦接(indirectly electrically coupled to)」(有居中的材料或元件於其間將各元件電性相連接)的含意。舉例而言,「在基材上方(over a substrate)」形成元件可包括直接在基材上形成元件而其間無居中的材料/元件存在,以及間接在基材上形成元件而其間有一或多個居中的材料/元件存在。

【符號說明】

- 10...記憶體單元
- 12...半導體基材/基材
- 14...源極區域/源極
- 14a...源極線
- 16...汲極區域/汲極
- 16a...位元線
- 16a1...第一位元線
- 16a2...第二位元線
- 18...通道區域
- 20...浮閘
- 22...控制閘

22a...第一部分/控制閘線/控制閘
 22a1...控制閘線
 22a2...控制閘線
 22b...第二部分
 24...中間絕緣體
 26...閘極氧化物
 28...選擇閘
 28a...選擇閘線/輸入線/線
 28a1...選擇閘線/選擇線/線
 28a2...選擇閘線/線
 30...抹除閘
 30a...抹除閘線
 31...數位類比轉換器
 32...VMM（向量矩陣乘法）
 32a...輸入 VMM
 32b...VMM
 33...記憶體陣列/非揮發性記憶體
 34...抹除閘及字線閘解碼器
 35...控制閘解碼器
 36...位元線解碼器
 37...源極線解碼器
 38...加總運算放大器
 39...啟動功能電路
 40...數位類比轉換器(DAC)
 42...電壓至電流轉換器 V/I Conv
 44...電流比較器 IComp
 46...電壓比較器 VComp
 48...電流至電壓轉換器 I/V Conv
 50...電流至電壓對數轉換器

52...電壓至電流對數轉換器
 54...參考 Gnd 的電流加算器
 56...參考 Vdd 的電流加算器
 3500...VMM
 3501...參考陣列
 3502...參考陣列
 3503...記憶體陣列
 3700...VMM
 3701...參考陣列
 3702...參考陣列
 3703...記憶體陣列
 3900...VMM
 3901...參考陣列
 3902...參考陣列
 3903...記憶體陣列
 4100...VMM
 4101...參考陣列
 4102...參考陣列
 4103...記憶體陣列
 4300...調諧演算法
 4301...步驟
 4302...步驟
 4303...步驟
 4304...步驟
 4305...步驟
 4306...步驟
 4307...步驟
 4308...步驟
 4309...步驟

4310...步驟
 4311...步驟
 4320...查找表
 4400...調諧演算法
 4401...步驟
 4402...步驟
 4403...步驟
 4404...步驟
 4405...步驟
 4406...步驟
 4407...步驟
 4408...步驟
 4500...調諧演算法
 4501...步驟
 4502...步驟
 4503...步驟
 4504...步驟
 4505...步驟
 4506...步驟
 4507...步驟
 4508...步驟
 4509...步驟
 4510...粗略演算法
 4511...精細演算法
 4600...均勻步階演算法
 4700...均勻對數步階（除以十）、粗略/精細演算法
 4800...二元搜尋步階、粗略/精細演算法

【發明申請專利範圍】

【請求項1】 一種用以在非揮發性記憶體單元陣列內識別可被迅速程式化的非揮發性記憶體單元之方法，該方法包含：

抹除一非揮發性記憶體單元；

以一第一電壓程式化該非揮發性記憶體單元；

測量在一讀取操作期間通過該非揮發性記憶體單元之一第一電流；

以一第二電壓程式化該非揮發性記憶體單元；

測量在一讀取操作期間通過該非揮發性記憶體單元之一第二電流；

以及

若該第二電流與該第一電流之間的差異超過一臨限值，將指示該非揮發性記憶體單元可被迅速程式化的資料儲存在一查找表中，該資料包含用於該非揮發性記憶體單元之一位址。

【請求項2】 如請求項 1 之方法，其中該資料包含一單一位元。

【請求項3】 如請求項 1 之方法，其中該記憶體單元係一分離 2 閘極快閃記憶體單元。

【請求項4】 如請求項 1 之方法，其中該記憶體單元係一分離 3 閘極快閃記憶體單元。

【請求項5】 如請求項 1 之方法，其中該記憶體單元係一分離 4 閘極快閃記憶體單元。

【請求項6】 如請求項 1 之方法，其中該記憶體單元在一次臨限區域中讀取。

【請求項7】 一種用以在非揮發性記憶體單元陣列內識別可被迅速程式化的非揮發性記憶體單元之方法，該方法包含：

程式化一非揮發性記憶體單元；

以一第一電壓抹除該非揮發性記憶體單元；

測量在一讀取操作期間通過該非揮發性記憶體單元之一第一電流；

以一第二電壓抹除該非揮發性記憶體單元；

測量在一讀取操作期間通過該非揮發性記憶體單元之一第二電流；

及

若該第二電流與該第一電流之間的差異超過一臨限值，將指示該非揮發性記憶體單元可被迅速程式化的資料儲存在一查找表中，該資料包含用於該非揮發性記憶體單元之一位址。

【請求項8】 如請求項 7 之方法，其中該資料包含一單一位元。

【請求項9】 如請求項 7 之方法，其中該記憶體單元係一分離 2 閘極快閃記憶體單元。

【請求項10】 如請求項 7 之方法，其中該記憶體單元係一分離 3 閘極快閃記憶體單元。

【請求項11】 如請求項 7 之方法，其中該記憶體單元係一分離 4 閘極快閃記憶體單元。

【請求項12】 如請求項 7 之方法，其中該記憶體單元係在一次臨限區域中讀取。

【請求項13】 一種用以基於非揮發性類比神經形態記憶體單元之一程式化速度特性而程式化複數個單元之方法，該方法包含：

藉由判定用於該單元之一位址是否被包含於一查找表中來判定單元之該程式化速度特性；

若該程式化速度特性具有一第一值，對該單元執行一第一調諧演算法以在該單元的一浮閘上達到一所欲的電荷位準；以及

若該程式化速度特性具有一第二值，對該單元執行一第二調諧演

算法以在該單元的該浮閘上達到該所欲的電荷位準。

【請求項14】 如請求項 13 之方法，其中該第一調諧演算法利用大於該第二調諧演算法所利用的程式化電壓增量之該等程式化電壓增量。

【請求項15】 如請求項 13 之方法，其中該第一調諧演算法利用大於該第二調諧演算法所利用的程式化脈衝寬度增量之該等程式化脈衝寬度增量。

【請求項16】 如請求項 13 之方法，其中該記憶體單元係一分離 2 閘極快閃記憶體單元。

【請求項17】 如請求項 13 之方法，其中該記憶體單元係一分離 3 閘極快閃記憶體單元。

【請求項18】 如請求項 13 之方法，其中該記憶體單元係一分離 4 閘極快閃記憶體單元。

【請求項19】 如請求項 13 之方法，其中該記憶體單元於一次臨限區域中操作。

【請求項20】 如請求項 13 之方法，其中該程式化速度特性係儲存在非揮發性記憶體中。

【請求項21】 如請求項 13 之方法，其中該程式化速度特性係儲存在一表中，該表包含具有該程式化速度特性之非揮發性記憶體單元的位址。

【請求項22】 如請求項 13 之方法，其中該程式化速度特性係儲存在一表中，該表包含用於該陣列中之該等非揮發性記憶體單元之各者的一單位元。

【請求項23】 如請求項 21 之方法，其進一步包含：當針對一非揮發性記憶體單元之該程式化速度特性改變時更新該表。

【請求項24】 如請求項 22 之方法，其進一步包含：當針對一非揮發性記憶體單元之該程式化速度特性改變時更新該表。

【請求項25】 如請求項 13 之方法，其中該等記憶體單元係配置在一陣列中，其中字線及源極線配置在一水平方向上，且位元線在一垂直方向上，其中該等位元線操作為電流輸出神經元。

【請求項26】 如請求項 25 之方法，其中將電壓輸入提供給該等控制閘。

【請求項27】 如請求項 26 之方法，其中使用流入至一經二極體連接調諧之參考單元的一輸入電流以提供該電壓輸入。

【請求項28】 如請求項 27 之方法，其中該經二極體連接調諧之參考單元係使用於該陣列中的一列記憶體單元。

【請求項29】 如請求項 13 之方法，其中該等記憶體單元係配置在一陣列中，其中字線、源極線、及控制閘線配置在一水平方向上，且抹除閘線及位元線配置在一垂直方向上，其中該等位元線操作為電流輸出神經元。

【請求項30】 如請求項 13 之方法，其中該等記憶體單元係配置在一陣列中，其中字線、源極線、及抹除閘線配置在一水平方向上，且控制閘線及位元線配置在一垂直方向上，其中該等源極線操作為電流輸出神經元。

【請求項31】 如請求項 13 之方法，其中該等記憶體單元係配置在一陣列中，其中字線及源極線配置在一水平方向上，且位元線配置在一垂直方向上，其中該等位元線操作為電流輸出神經元。

【請求項32】 如請求項 31 之方法，其中電壓輸入要至該等字線。

【請求項33】 如請求項 13 之方法，其中該等記憶體單元係配置在

一陣列中，其中源極線配置在一水平方向上，且字線及位元線組織在一垂直方向上，其中該等源極線操作為電流輸出神經元。

【請求項34】 如請求項 32 之方法，其中使用流入至一經二極體連接調諧之參考單元的一輸入電流以提供該電壓輸入。

【請求項35】 如請求項 34 之方法，其中該經二極體連接調諧之參考單元係使用於一列。

【請求項36】 一種用以程式化非揮發性類比神經形態記憶體單元之方法，該方法包含：

抹除該單元；

在該單元上執行一粗略程式化序列，其包含：

對該單元執行一第一程式化操作，其中每次該第一程式化操作被執行時，施加等於前次第一程式化操作中所使用之一程式化電壓加上一第一電壓增量之程式化電壓；及

重複該第一程式化操作，直到在一讀取操作期間通過該單元的電流超過一第一電流臨限；以及

在該單元上執行一精細程式化序列，其包含：

對該單元執行一第二程式化操作，其中每次該第一程式化操作被執行時，施加等於前次第二程式化操作中所使用之一程式化電壓加上一第二電壓增量之一程式化電壓，其中該第二電壓增量小於該第一電壓增量；以及

重複該第二程式化操作，直到在一讀取操作期間通過該單元的該電流超過一第二電流臨限。

【請求項37】 如請求項 36 之方法，其進一步包含：

若該第一程式化操作執行的一次數超過一臨限，則將該單元識別

為不欲使用的一不良單元。

【請求項38】 如請求項 36 之方法，其中該記憶體單元係一分離 2 閘極快閃記憶體單元。

【請求項39】 如請求項 36 之方法，其中該記憶體單元係一分離 3 閘極快閃記憶體單元。

【請求項40】 如請求項 36 之方法，其中該記憶體單元係一分離 4 閘極快閃記憶體單元。

【請求項41】 如請求項 36 之方法，其中該記憶體單元於一次臨限區域中操作。

【請求項42】 一種類神經網路裝置，包含：

一第一複數個突觸，其經組配以接收一第一複數個輸入並由其產生一第一複數個輸出，其中該等第一複數個突觸包含：

複數個記憶體單元，其中該等記憶體單元之各者包括於一半導體基材中形成之間隔開的源極區域及汲極區域，而一通道區域在該源極區域與該汲極區域之間延伸，設置於該通道區域之一第一部分上方並與該第一部分絕緣之一浮閘，及具有一第一部分的一第一閘極，該第一閘極之第一部分設置於該通道區域之一第二部分上方並與該第二部分絕緣；

該等複數個記憶體單元之各者係組配以儲存對應於該浮閘上之電子之數目的一權重值；

該等複數個記憶體單元係組配以基於該等第一複數個輸入及所儲存的該等權重值來產生該等第一複數個輸出；

其中該等第一複數個突觸之該等記憶體單元係配置成列與行，且其中該等第一複數個突觸包含：

複數條第一線，各條第一線將該等記憶體單元之該等列之一者中的該等第一閘極電連接在一起；

複數條第二線，各條第二線將該等記憶體單元之該等列之一者中的該等源極區域電連接在一起；

複數條第三線，各條第三線將該等記憶體單元之該等行之一者中的該等汲極區域電連接在一起；

其中該等第一複數個突觸係組配以接收該等第一複數個輸入作為該等複數條第二線上的電壓，以及提供該等第一複數個輸出作為該等複數條第三線上的電流。

【請求項43】 如請求項 42 之類神經網路裝置，其中針對該等複數個記憶體單元之各者，該第一閘極包括向上延伸且延伸於該浮閘上方的一第二部分。

【請求項44】 如請求項 42 之類神經網路裝置，進一步包含：

一第一複數個神經元，其經組配以接收該等第一複數個輸出。

【請求項45】 如請求項 44 之類神經網路裝置，進一步包含：

一第二複數個突觸，其經組配以從該等第一複數個神經元接收一第二複數個輸入並由其產生一第二複數個輸出，其中該等第二複數個突觸包含：

複數個第二記憶體單元，其中該等第二記憶體單元之各者包括於該半導體基材中形成之間隔開的第二源極區域及第二汲極區域，而一第二通道區域在該第二源極區域與該第二汲極區域之間延伸，設置於該第二通道區域之一第一部分上方並與該第一部分絕緣之一第二浮閘，及具有一第一部分的一第二閘極，該第二閘極之第一部分設置於該第二通道區域之一第二部分上方並與該第

二部分絕緣；

該等複數個第二記憶體單元之各者係組配以儲存對應於該第二浮閘上之電子之數目的一第二權重值；

該等複數個第二記憶體單元係組配以基於該等第二複數個輸入及所儲存的該等第二權重值來產生該等第二複數個輸出；

其中該等第二複數個突觸之該等第二記憶體單元係配置成列與行，且其中該等第二複數個突觸包含：

複數條第四線，各條第四線將該等第二記憶體單元之該等列之一者中的該等第二閘極電連接在一起；

複數條第五線，各條第五線將該等第二記憶體單元之該等列之一者中的該等第二源極區域電連接在一起；

複數條第六線，各條第六線將該等第二記憶體單元之該等行之一者中的該等第二汲極區域電連接在一起；

其中該等第二複數個突觸係組配以接收該等第二複數個輸入作為該等複數條第五線上的電壓，以及提供該等第二複數個輸出作為該等複數條第六線上的電流。

【請求項46】 如請求項 45 之類神經網路裝置，其中針對該等複數個第二記憶體單元之各者，該第二閘極包括向上延伸且延伸於該第二浮閘上方的一第二部分。

【請求項47】 如請求項 45 之類神經網路裝置，進一步包含：

一第二複數個神經元，其經組配以接收該等第二複數個輸出。

【請求項48】 一種類神經網路裝置，包含：

一第一複數個突觸，其經組配以接收一第一複數個輸入並由其產生一第一複數個輸出，其中該等第一複數個突觸包含：

複數個記憶體單元，其中該等記憶體單元之各者包括於一半導體基材中形成之間隔開的源極區域及汲極區域，而一通道區域在該源極區域與該汲極區域之間延伸，設置於該通道區域之一第一部分上方並與該第一部分絕緣之一浮閘，設置於該通道區域之一第二部分上方並與該第二部分絕緣的一第一閘極，設置於該浮閘上方並與該浮閘絕緣的一第二閘極，及設置於該源極區域上方並與該源極區域絕緣的一第三閘極；

該等複數個記憶體單元之各者係組配以儲存對應於該浮閘上之電子之數目的一權重值；

該等複數個記憶體單元係組配以基於該等第一複數個輸入及所儲存的該等權重值來產生該等第一複數個輸出；

其中該等第一複數個突觸之該等記憶體單元係配置成列與行，且其中該等第一複數個突觸包含：

複數條第一線，各條第一線將該等記憶體單元之該等列之一者中的該等第一閘極電連接在一起；

複數條第二線，各條第二線將該等記憶體單元之該等列之一者中的該等第二閘極電連接在一起；

複數條第三線，各條第三線將該等記憶體單元之該等列之一者中的該等第三閘極電連接在一起；

複數條第四線，各條第四線將該等記憶體單元之該等列之一者中的該等源極區域電連接在一起；

複數條第五線，各條第五線將該等記憶體單元之該等行之一者中的該等汲極區域電連接在一起；

其中該等第一複數個突觸係組配以接收該等第一複數個輸入作為

該等複數條第二線上或該等複數條第三線上或該等複數條第四線上的電壓，以及提供該等第一複數個輸出作為該等複數條第五線上的電流。

【請求項49】 如請求項 48 之類神經網路裝置，其中該等第一複數個突觸係組配以接收該等第一複數個輸入作為該等複數條第二線上的電壓。

【請求項50】 如請求項 48 之類神經網路裝置，其中該等第一複數個突觸係組配以接收該等第一複數個輸入作為該等複數條第三線上的電壓。

【請求項51】 如請求項 48 之類神經網路裝置，其中該等第一複數個突觸係組配以接收該等第一複數個輸入作為該等複數條第四線上的電壓。

【請求項52】 如請求項 48 之類神經網路裝置，進一步包含：

一第一複數個神經元，其經組配以接收該等第一複數個輸出。

【請求項53】 如請求項 52 之類神經網路裝置，進一步包含：

一第二複數個突觸，其經組配以從該等第一複數個神經元接收一第二複數個輸入並由其產生一第二複數個輸出，其中該等第二複數個突觸包含：

複數個第二記憶體單元，其中該等第二記憶體單元之各者包括於該半導體基材中形成之間隔開的第二源極區域及第二汲極區域，而一第二通道區域在該第二源極區域與該第二汲極區域之間延伸，設置於該第二通道區域之一第一部分上方並與該第一部分絕緣之一第二浮閘，設置於該第二通道區域之一第二部分上方並與該第二部分絕緣之一第四閘極，設置於該第二浮閘上方並與該第二浮閘絕緣之一第五閘極，及設置於該第二源極區域上方並與

該第二源極區域絕緣之一第六閘極；

該等複數個第二記憶體單元之各者係組配以儲存對應於該第二浮閘上之電子之數目之一第二權重值；

該等複數個第二記憶體單元係組配以基於該等第二複數個輸入及所儲存的該等第二權重值來產生該等第二複數個輸出；

其中該等第二複數個突觸之該等第二記憶體單元係配置成列與行，且其中該等第二複數個突觸包含：

複數條第六線，各條第六線將該等第二記憶體單元之該等列之一者中的該等第四閘極電連接在一起；

複數條第七線，各條第七線將該等第二記憶體單元之該等列之一者中的該等第五閘極電連接在一起；

複數條第八線，各條第八線將該等第二記憶體單元之該等列之一者中的該等第六閘極電連接在一起；

複數條第九線，各條第九線將該等第二記憶體單元之該等列之一者中的該等第二源極區域電連接在一起；

複數條第十線，各條第十線將該等第二記憶體單元之該等行之一者中的該等第二汲極區域電連接在一起；

其中該等第二複數個突觸係組配以接收該等第二複數個輸入作為該等複數條第七線上或該等複數條第八線上或該等複數條第九線上的電壓，以及提供該等第二複數個輸出作為該等複數條第十線上的電流。

【請求項54】 如請求項 53 之類神經網路裝置，其中該等第二複數個突觸係組配以接收該等第二複數個輸入作為該等複數條第七線上的電壓。

【請求項55】 如請求項 53 之類神經網路裝置，其中該等第二複數

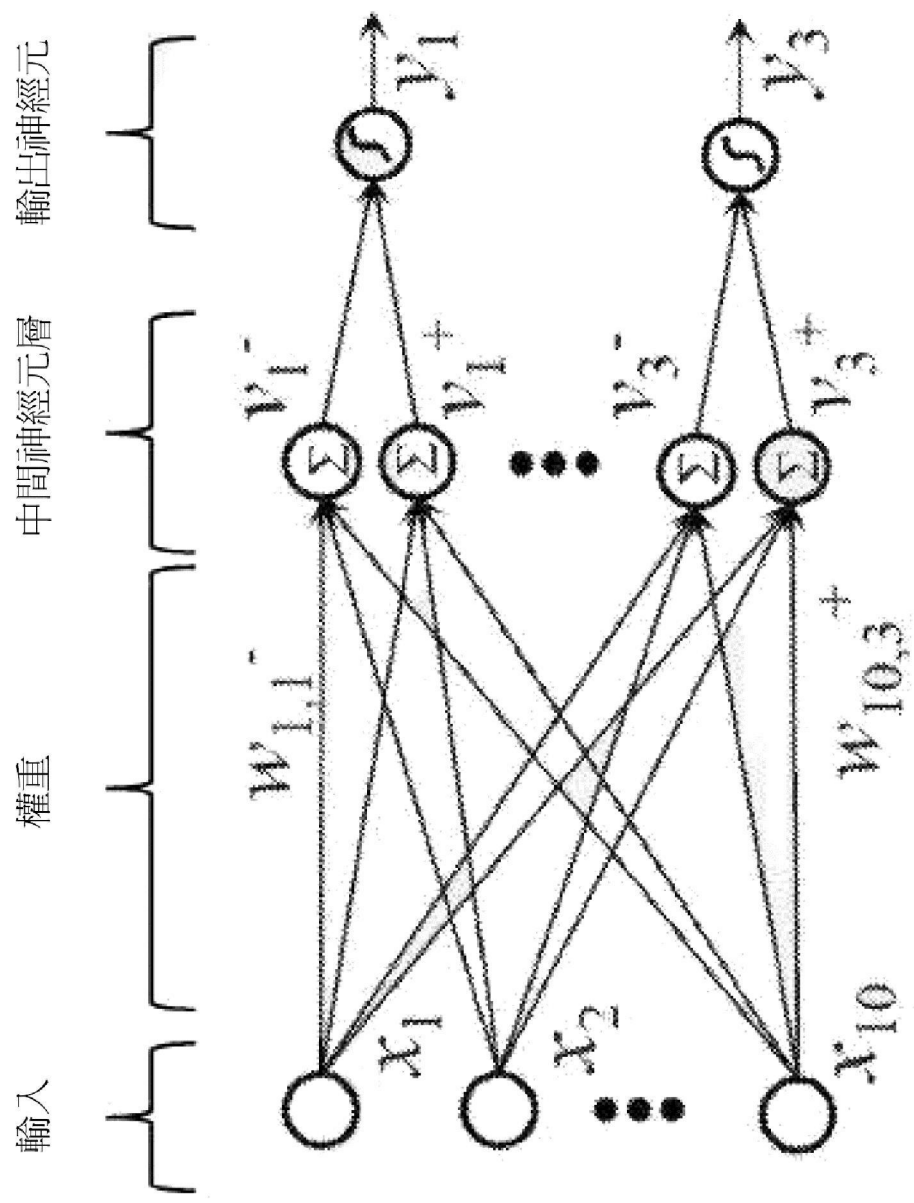
個突觸係組配以接收該等第二複數個輸入作為該等複數條第八線上的電壓。

【請求項56】 如請求項 53 之類神經網路裝置，其中該等第二複數個突觸係組配以接收該等第二複數個輸入作為該等複數條第九線上的電壓。

【請求項57】 如請求項 53 之類神經網路裝置，進一步包含：

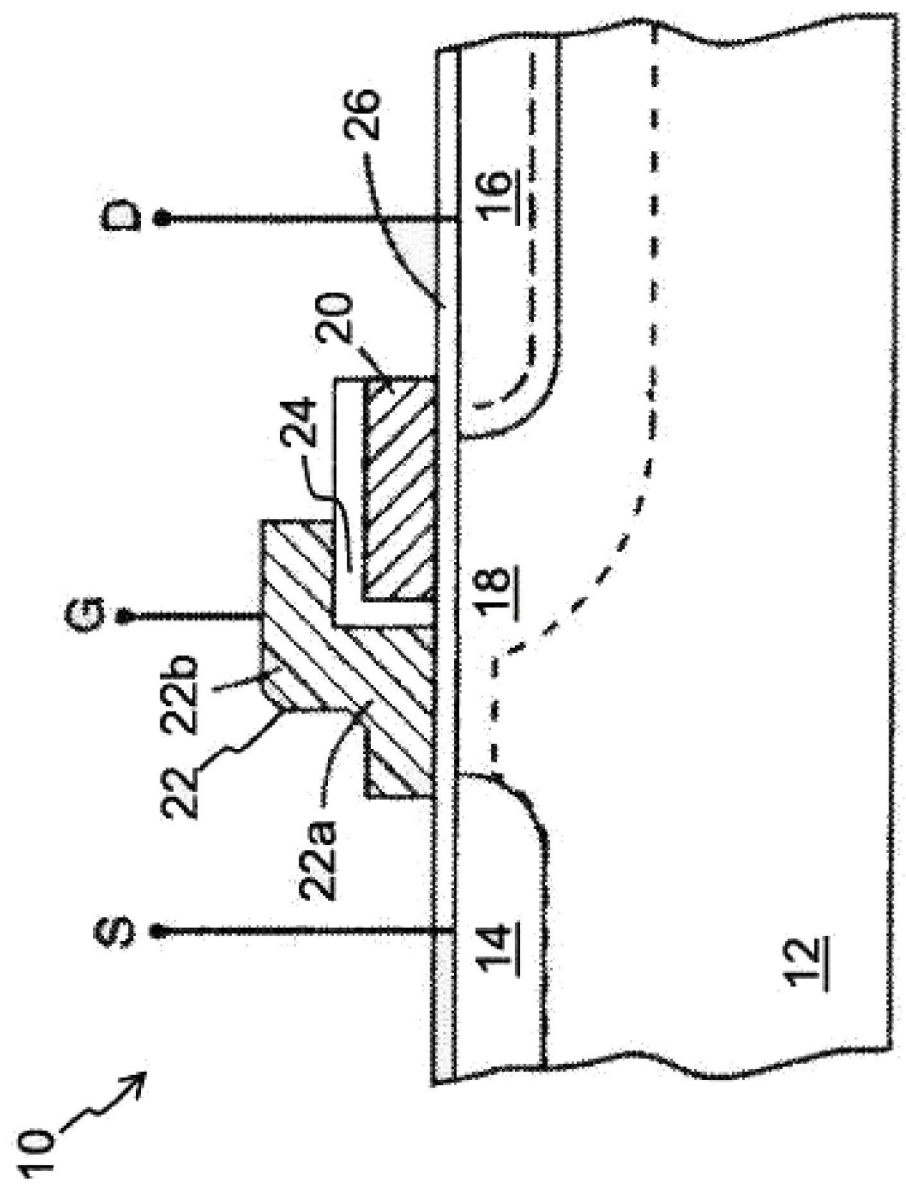
一第二複數個神經元，其經組配以接收該等第二複數個輸出。

【發明圖式】
圖式



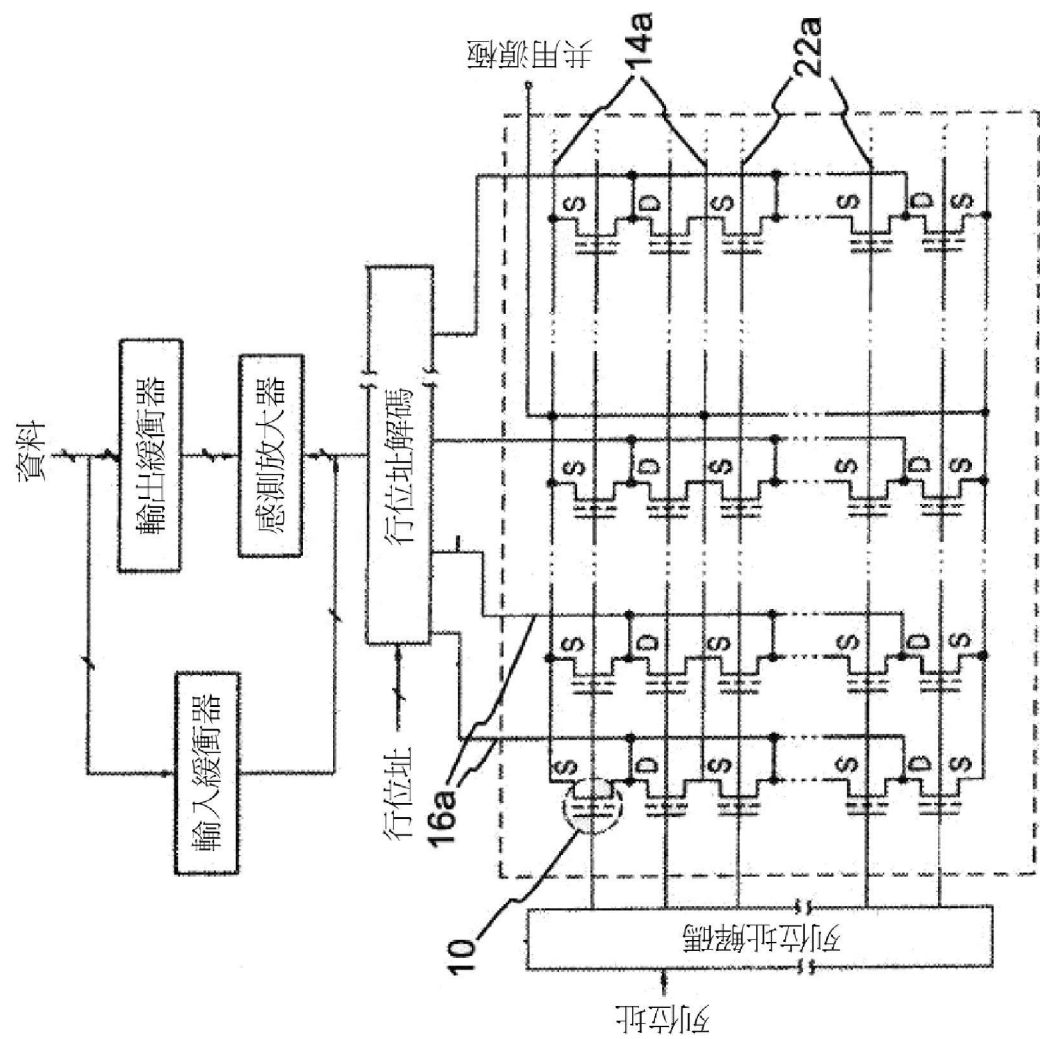
(先前技術)

【圖1】



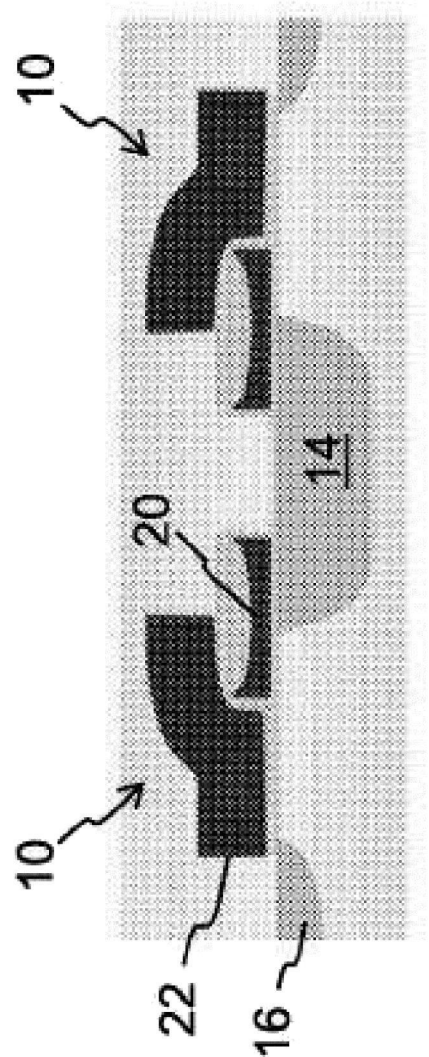
(先前技術)

【圖2】



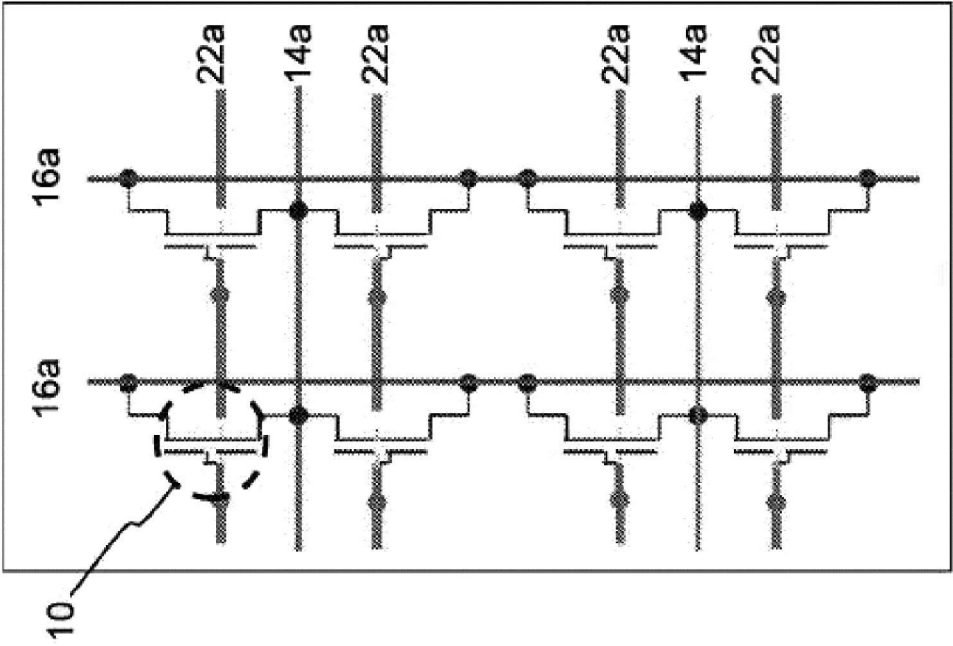
(先前技術)

【圖3】



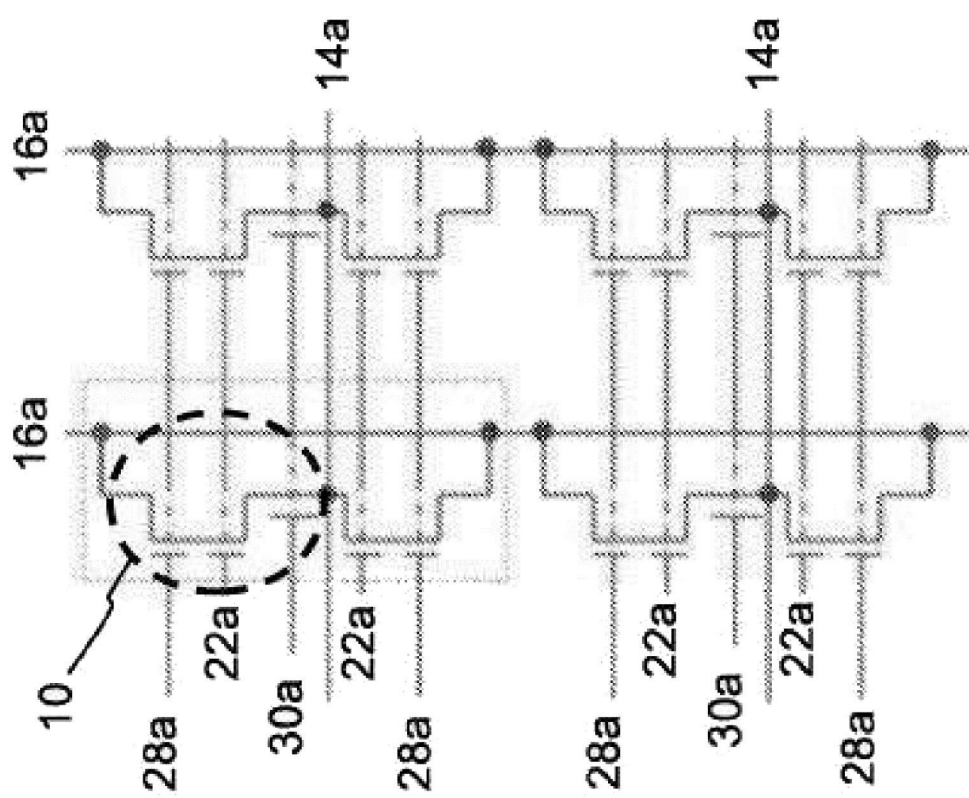
(先前技術)

【圖4】



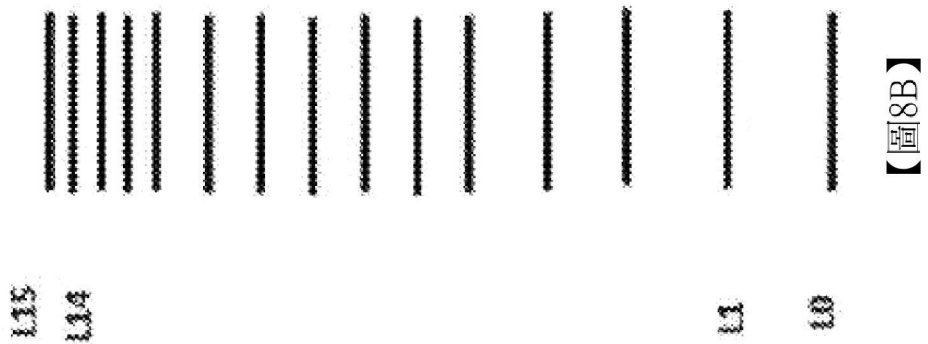
(先前技術)

【圖5】

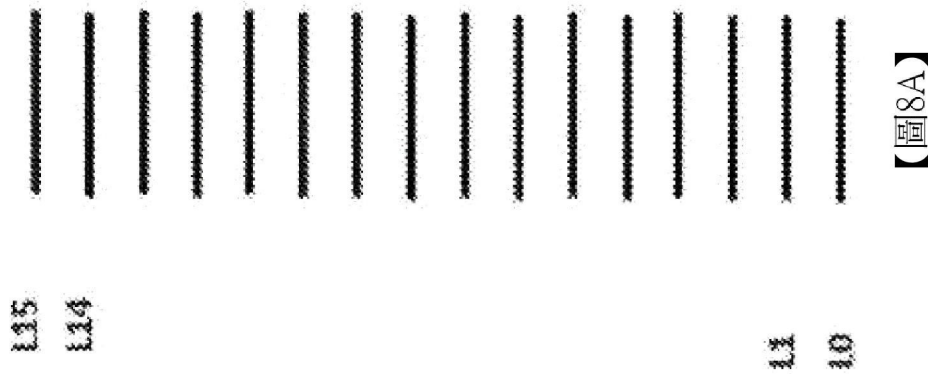


(先前技術)

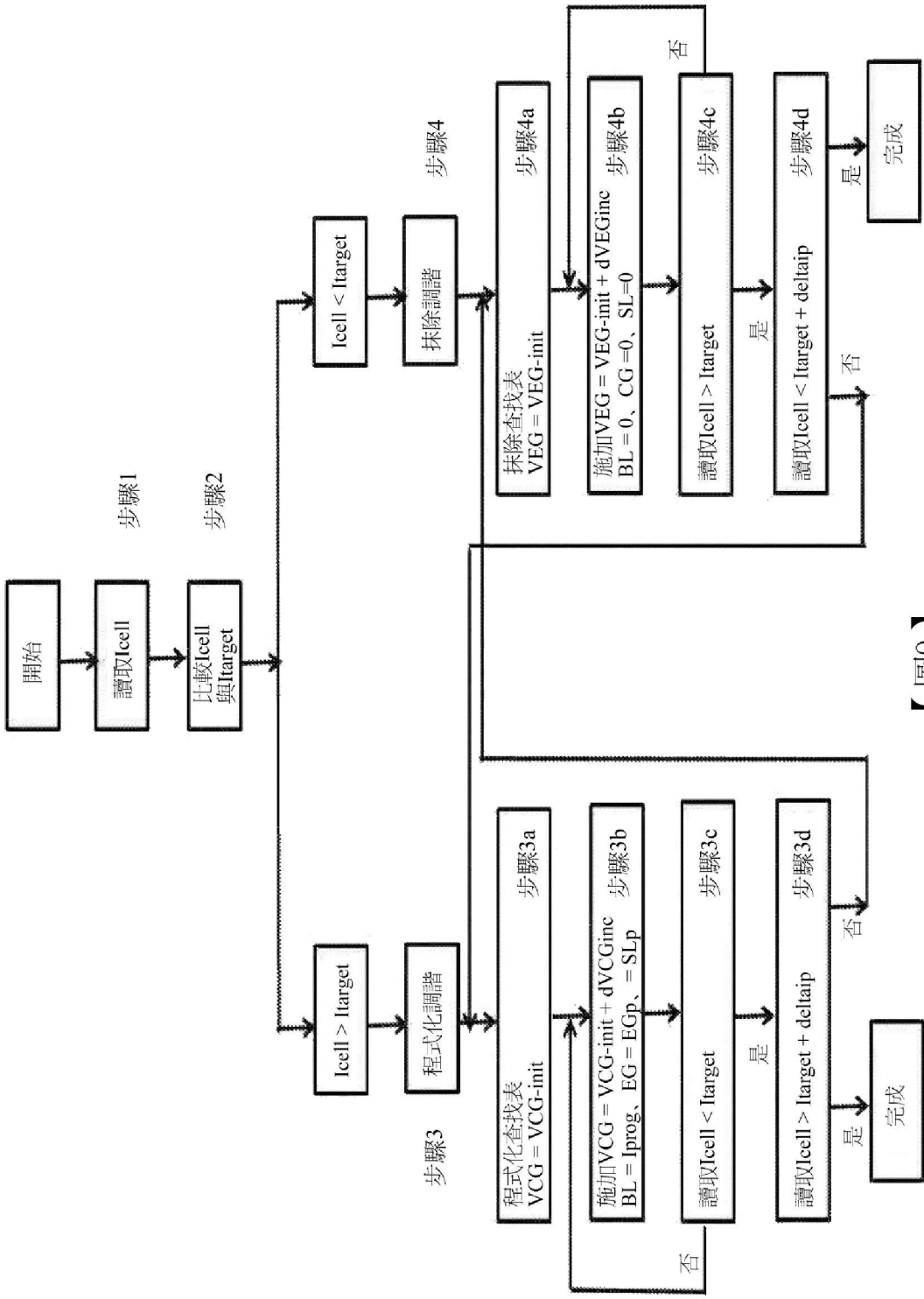
【圖7】



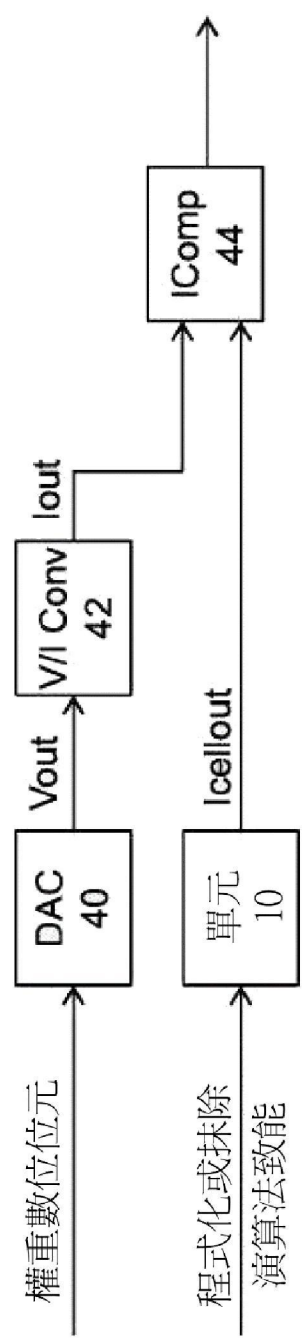
【圖8B】



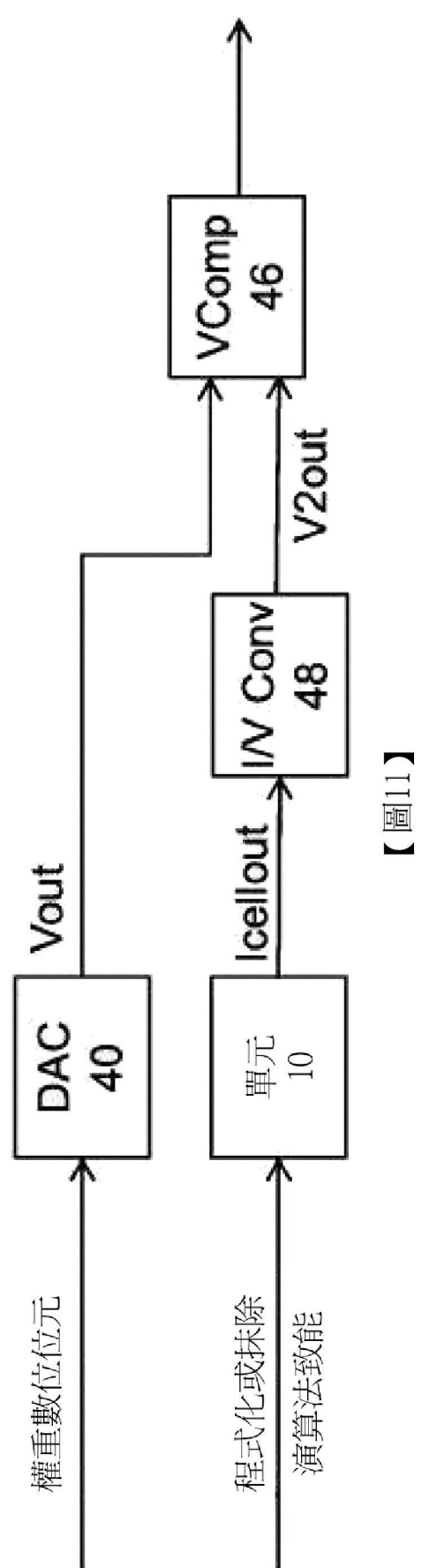
【圖8A】



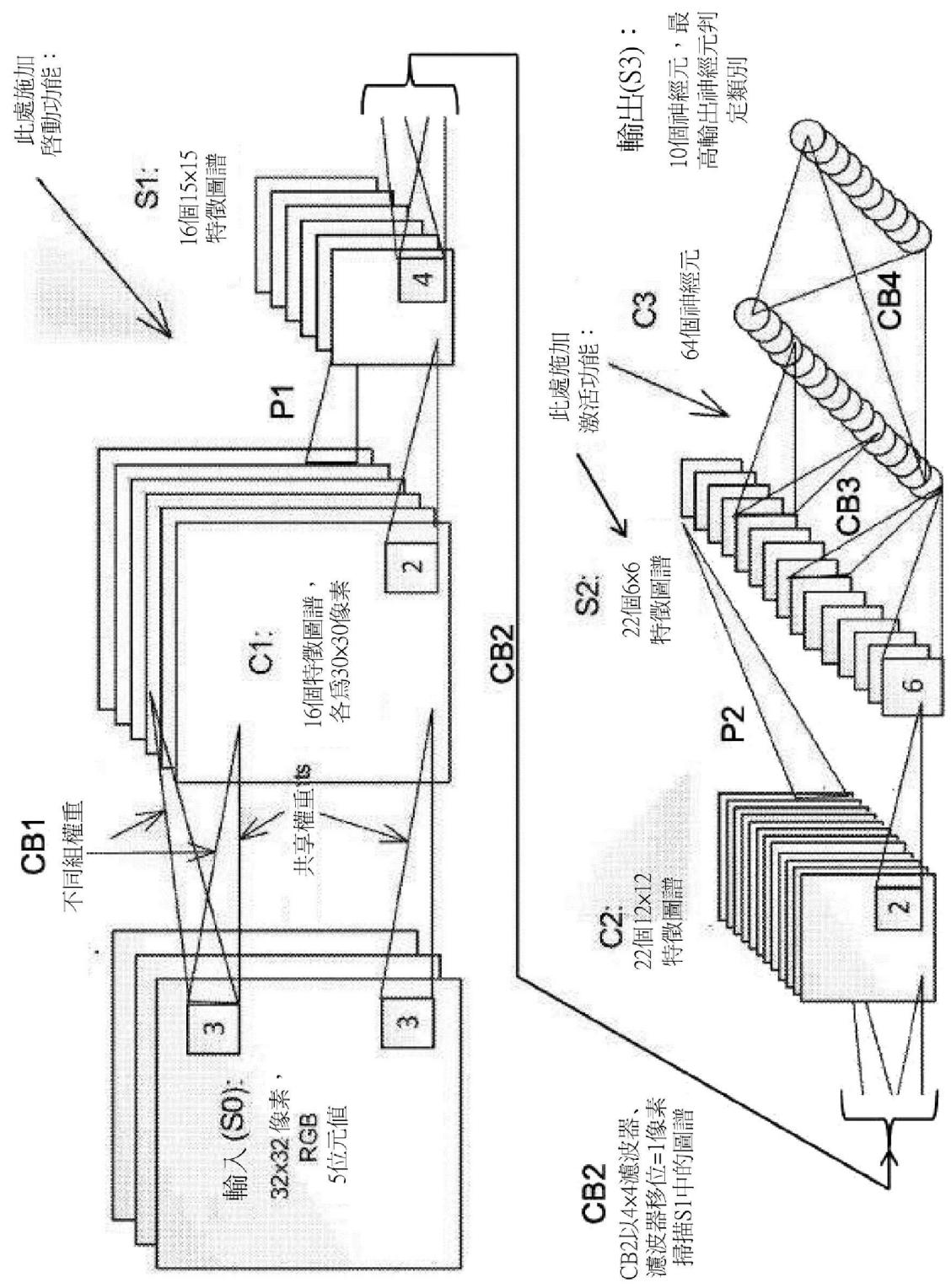
【圖9】



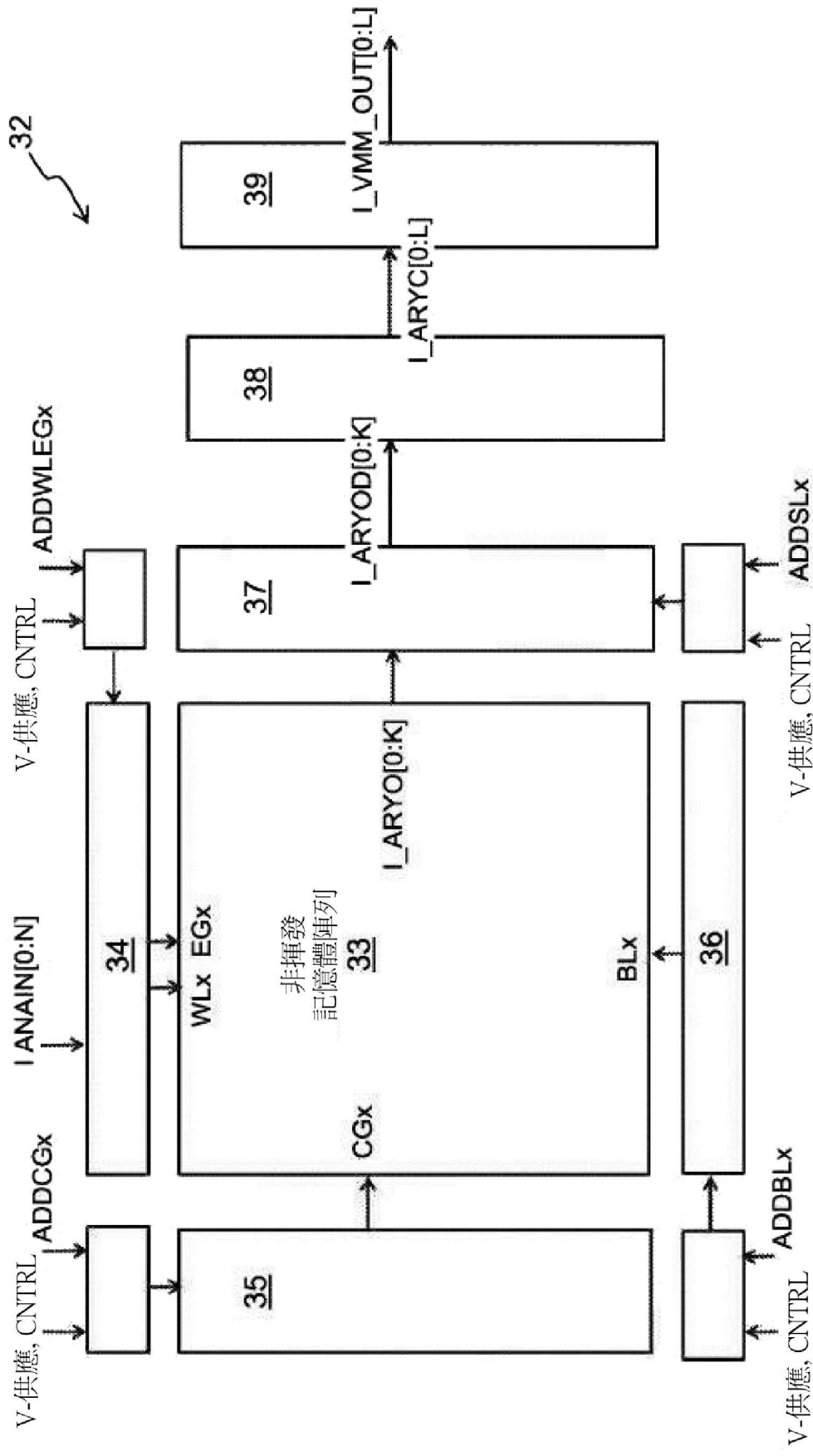
【圖10】



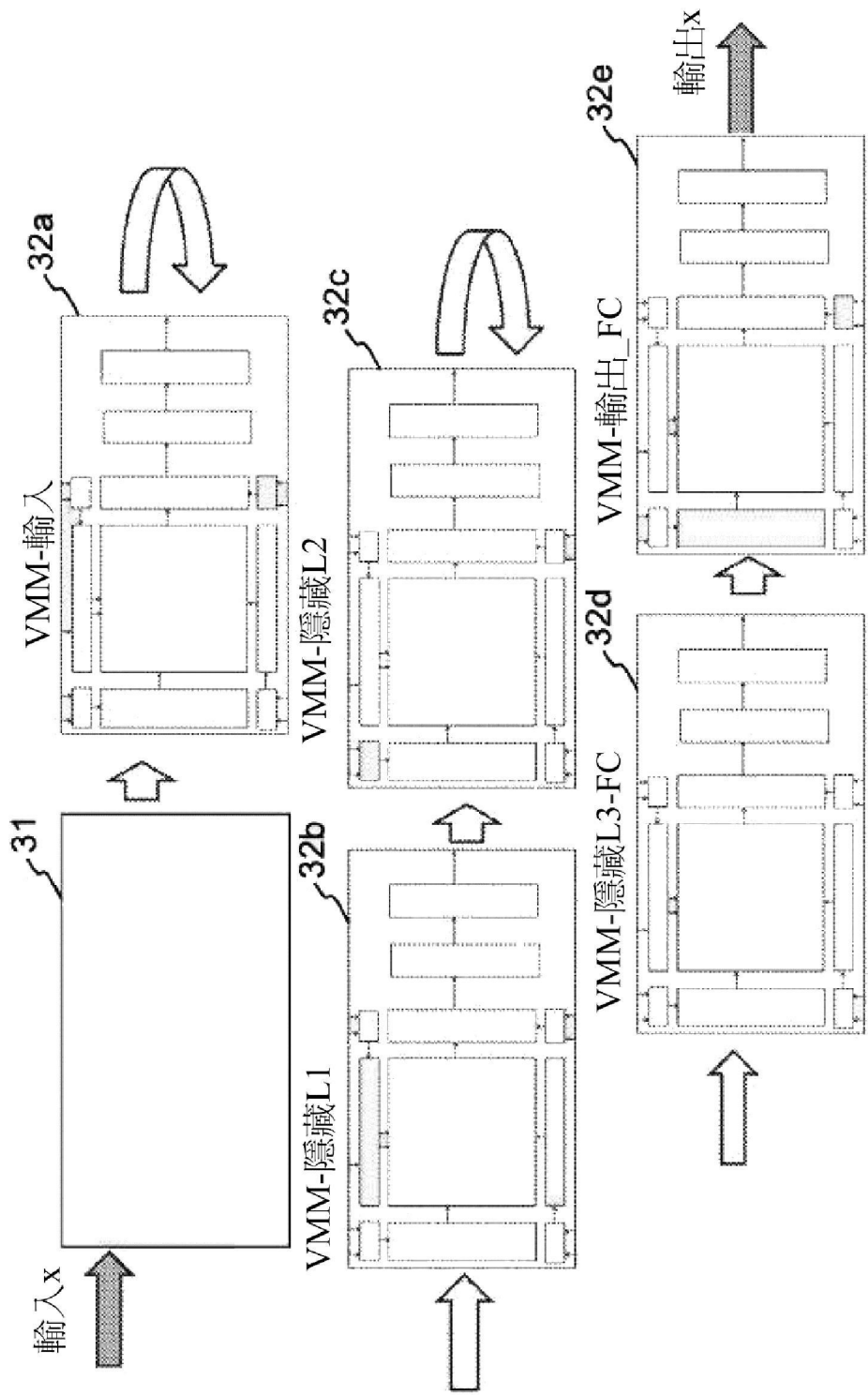
【圖11】



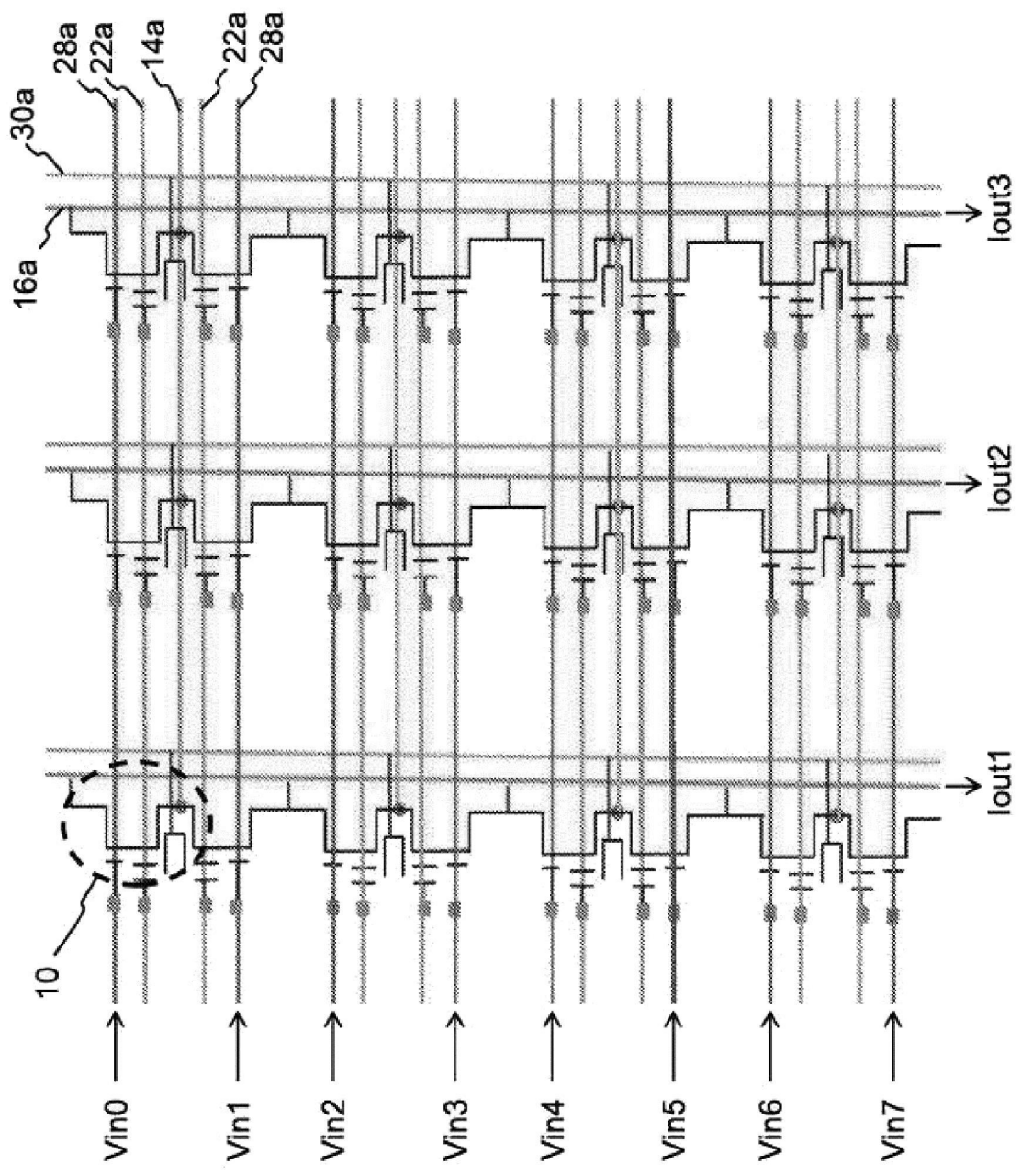
【圖12】



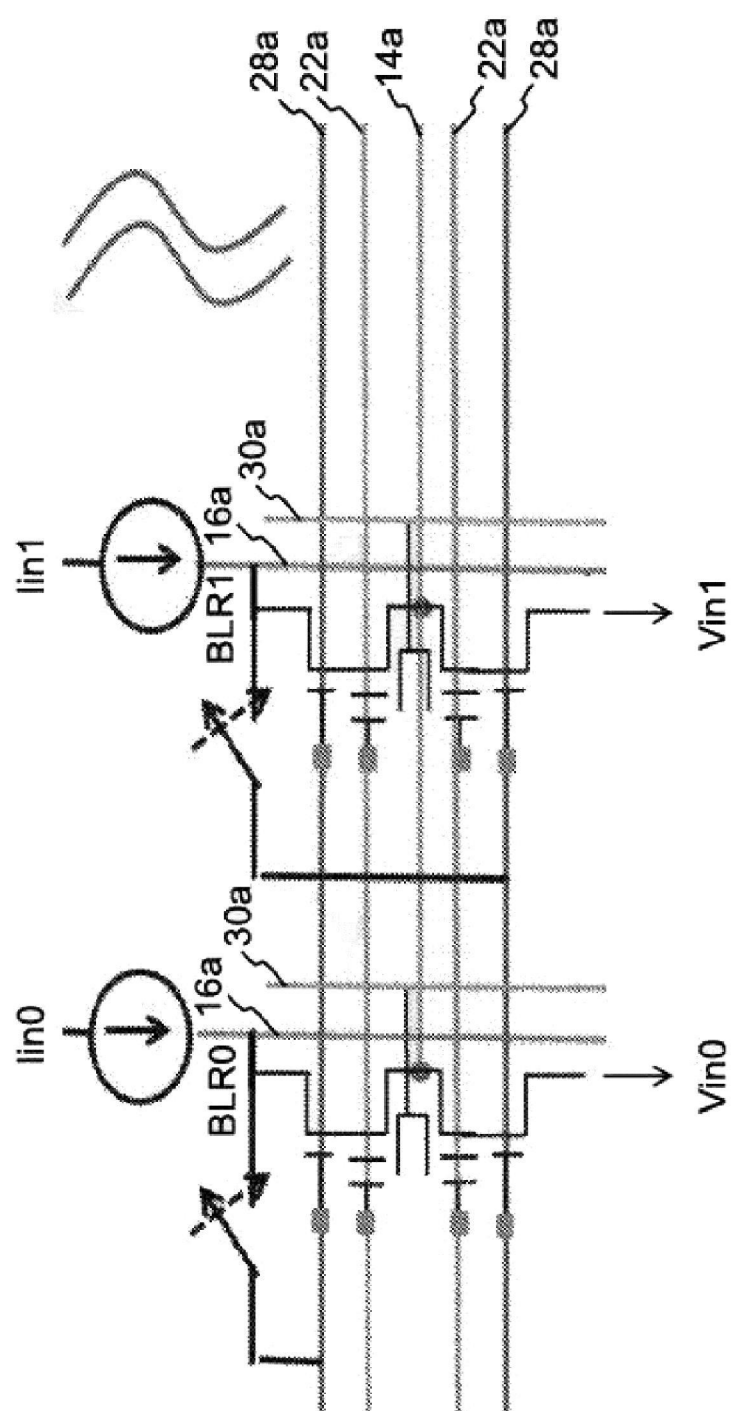
【圖13】



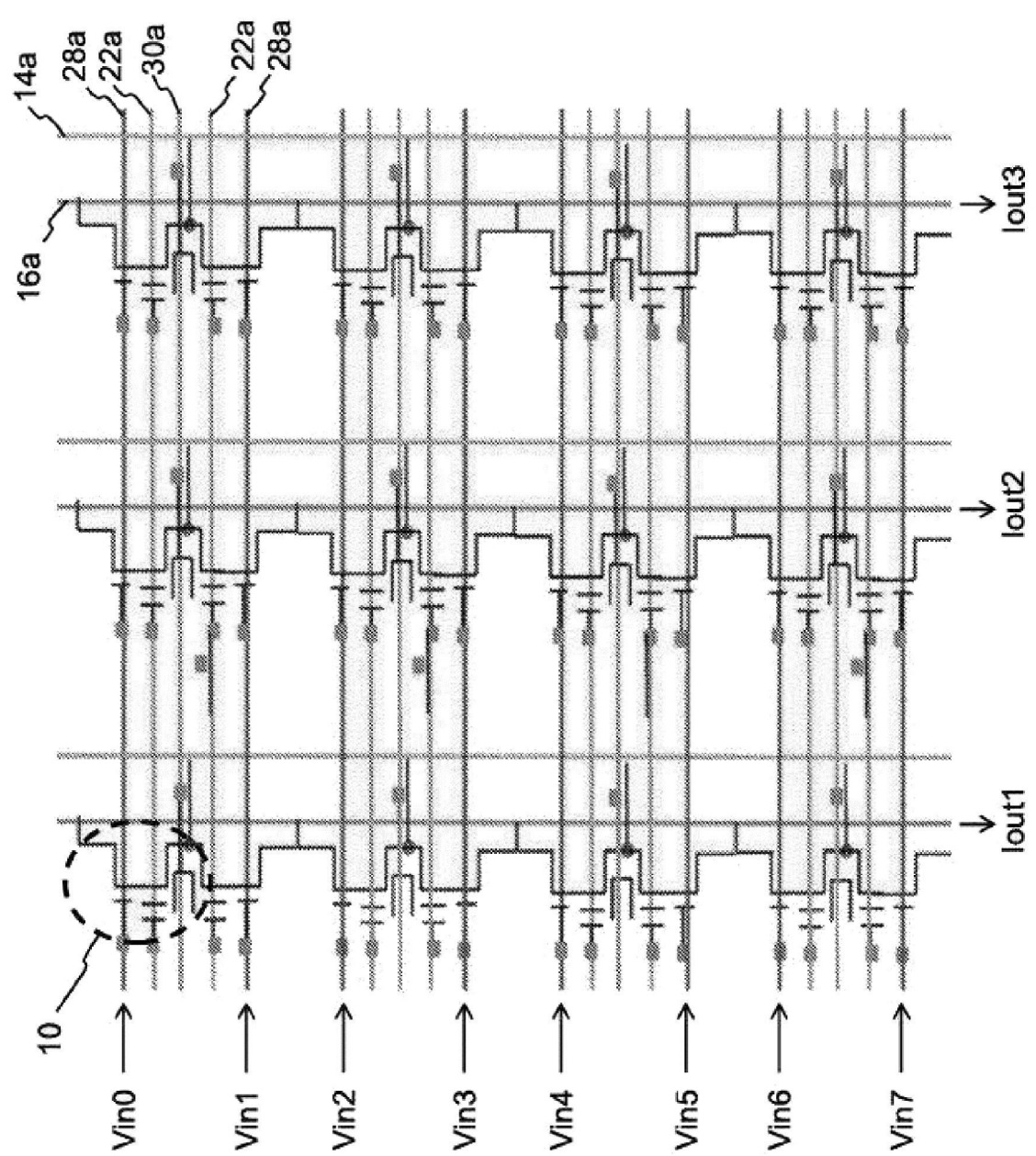
【圖14】



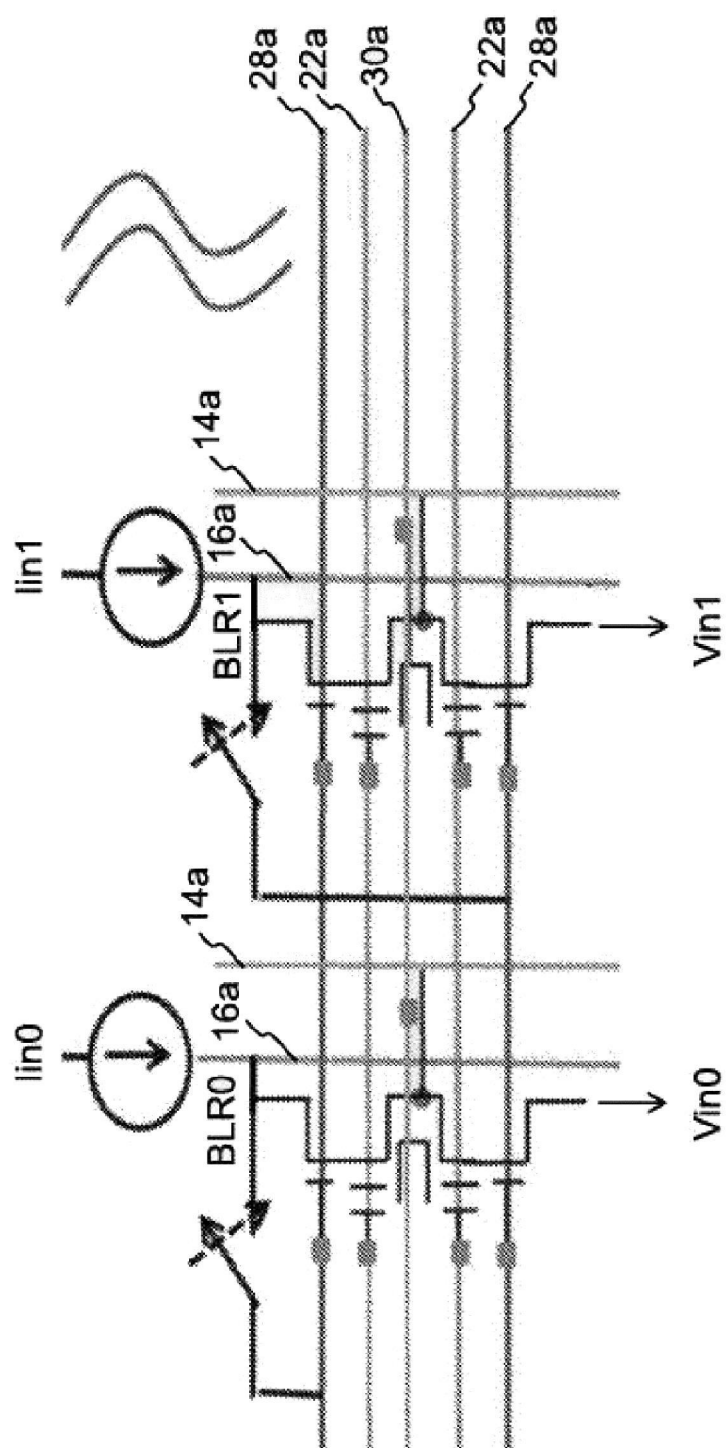
【圖15】



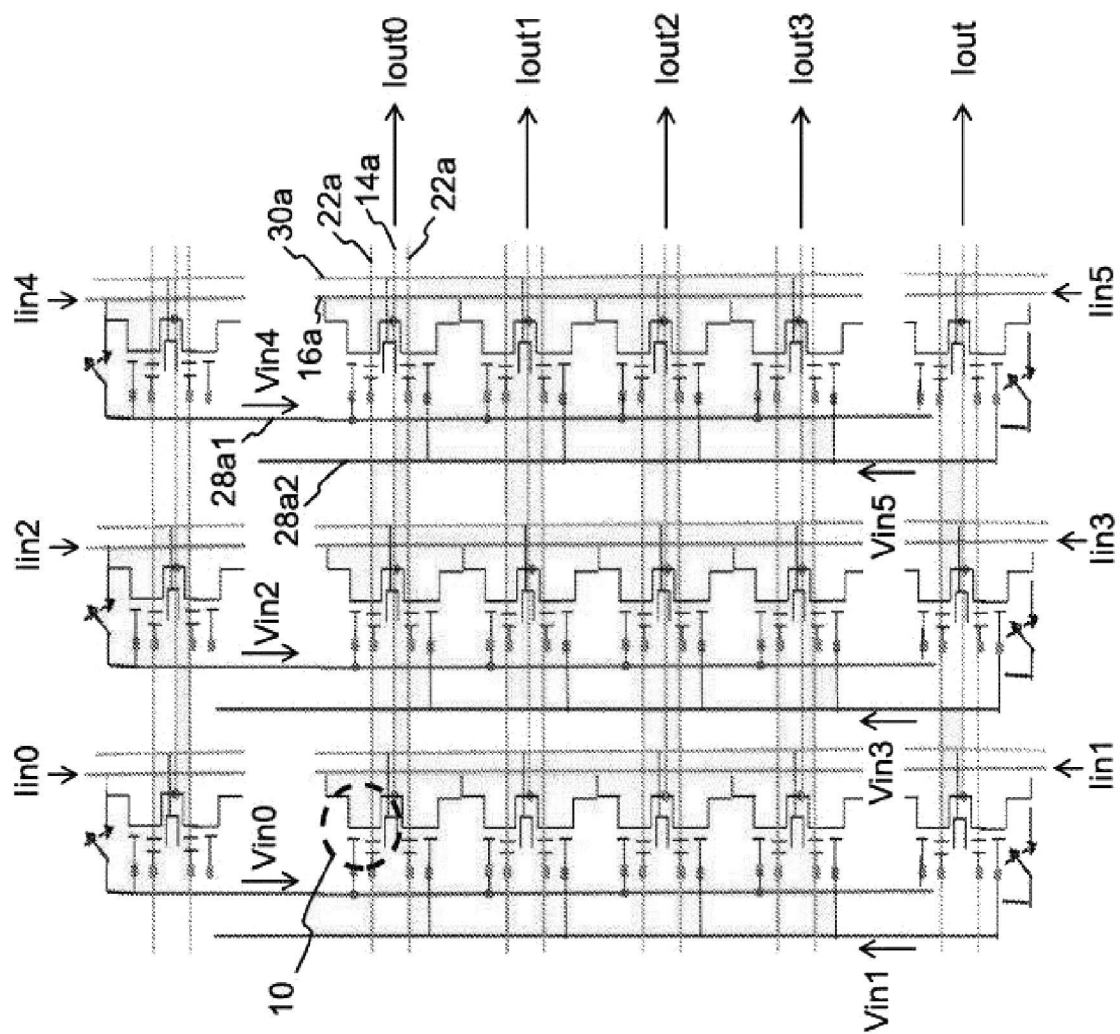
【16】



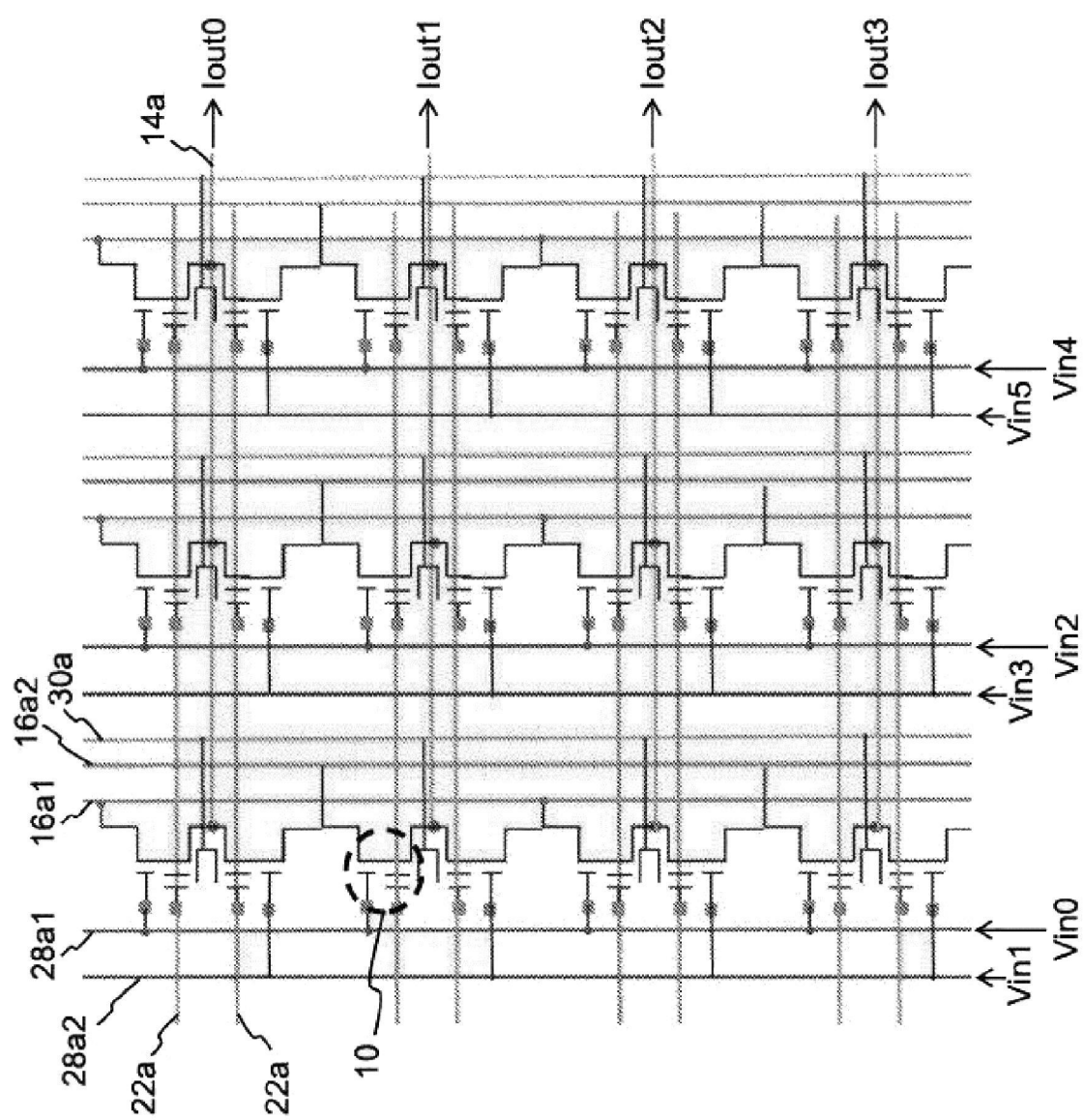
【圖17】



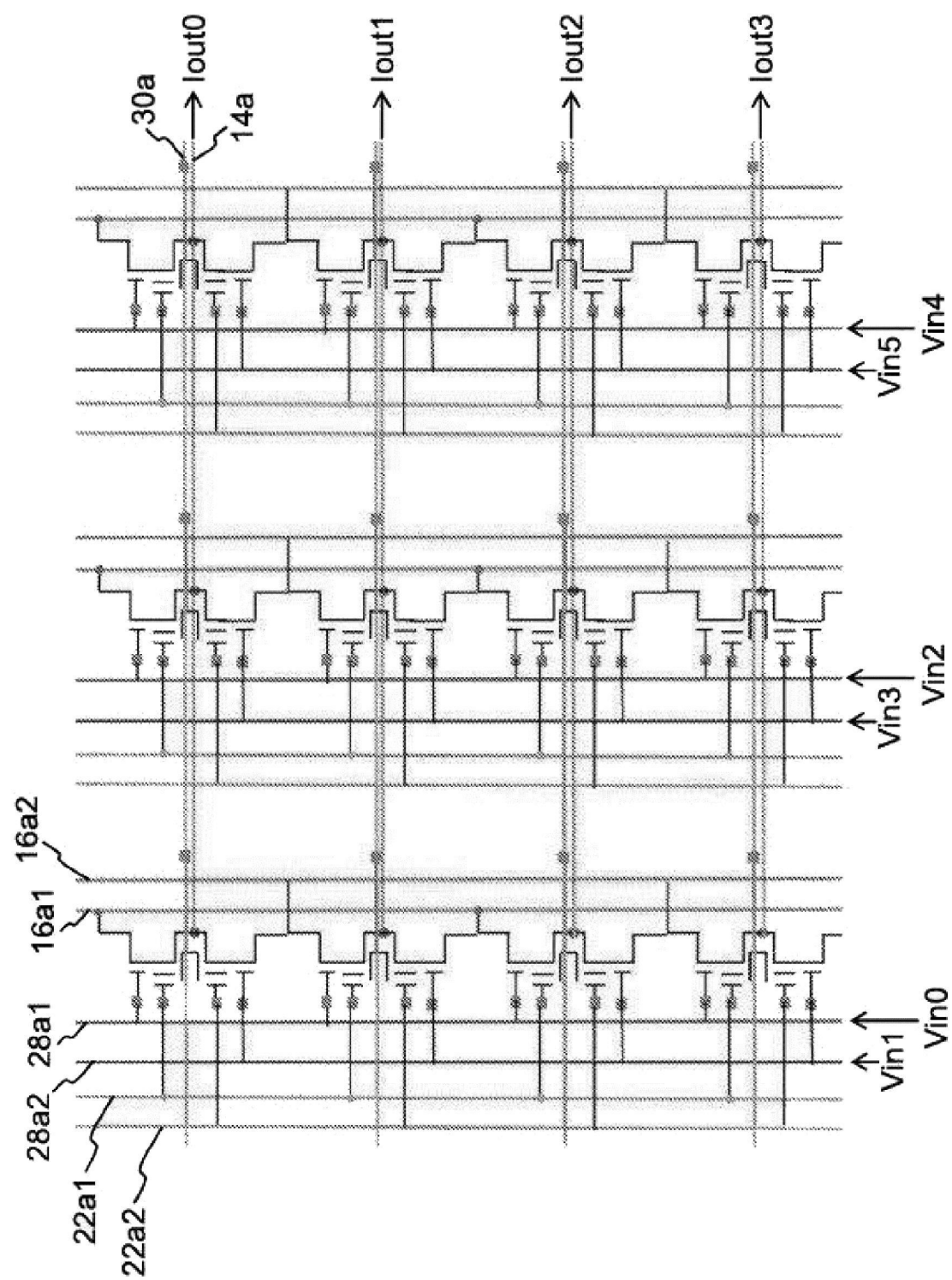
【18回】



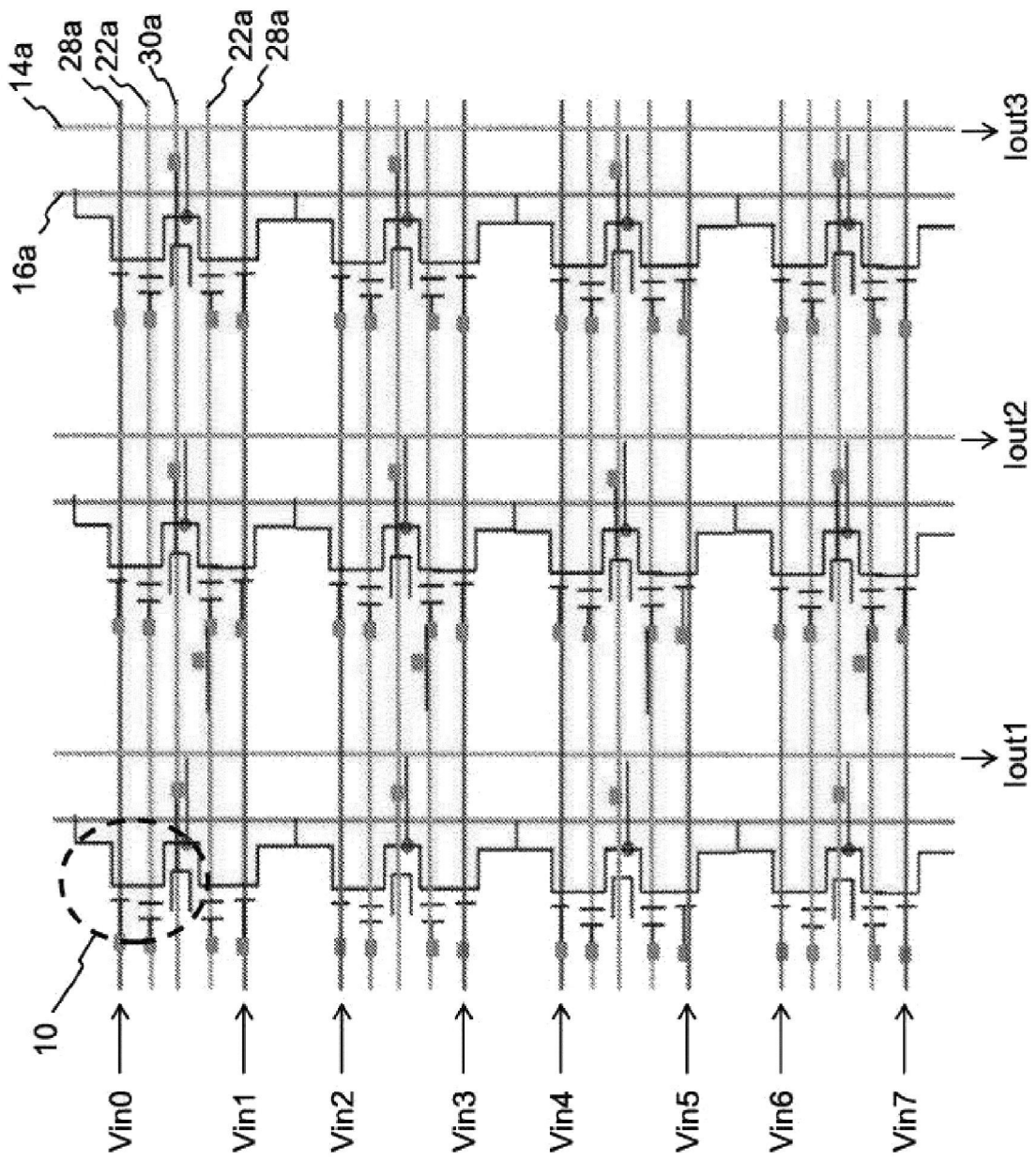
【圖19】



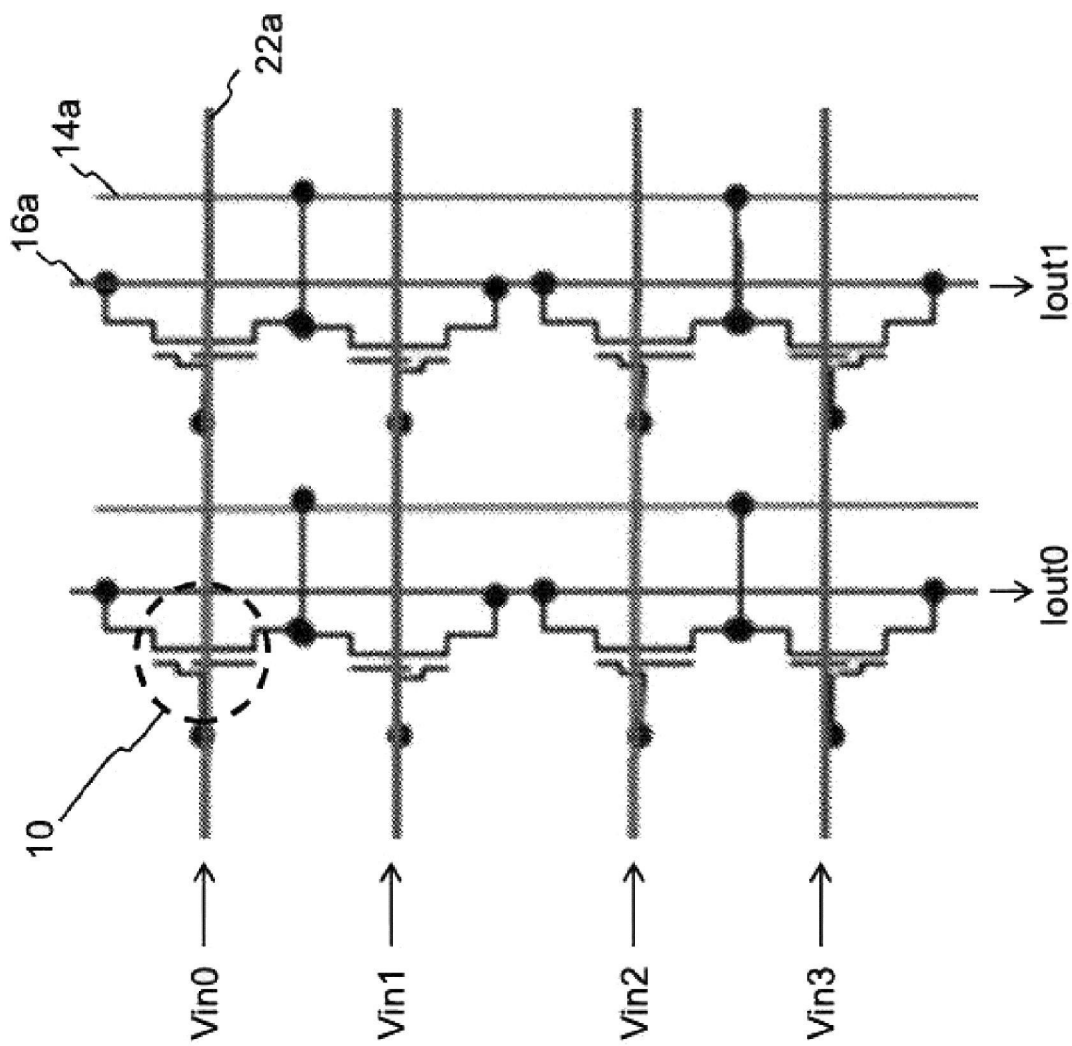
【圖20】



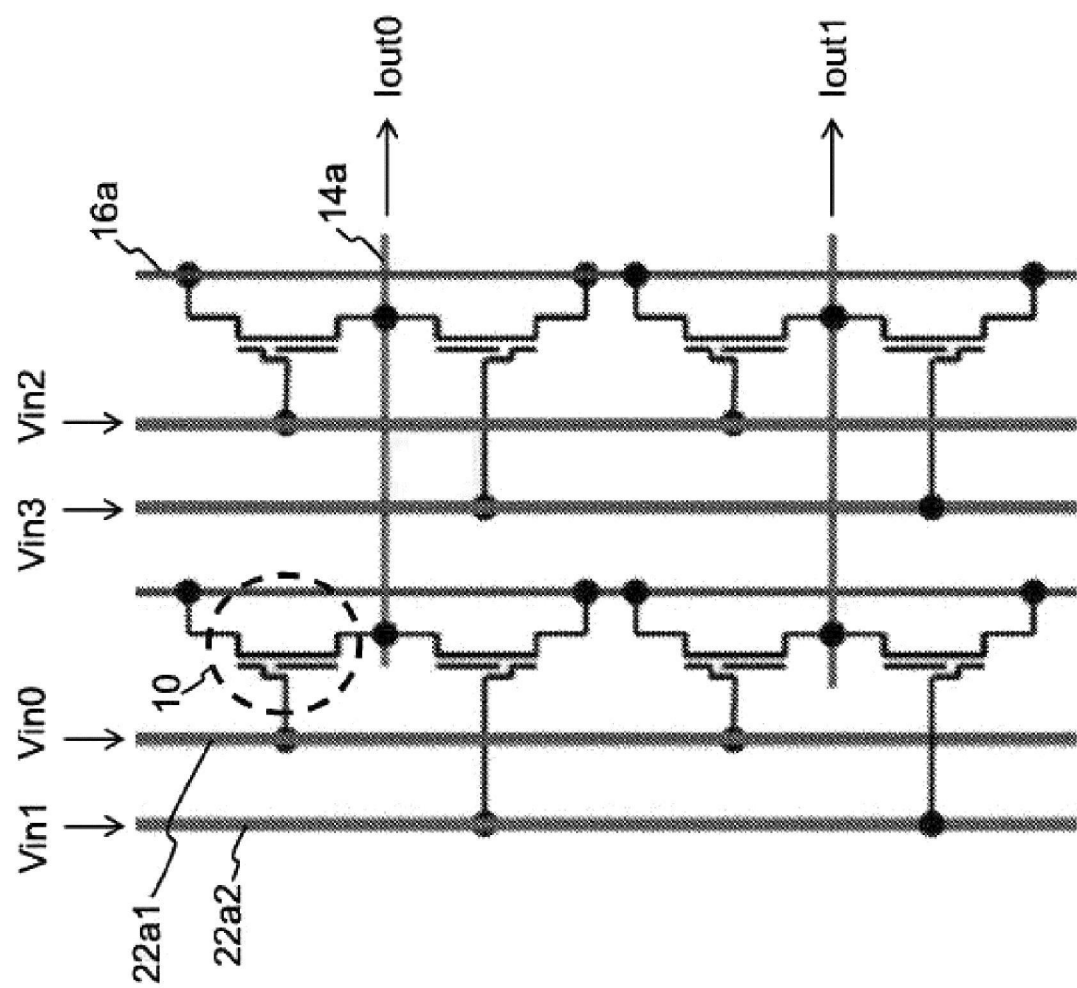
【圖21】



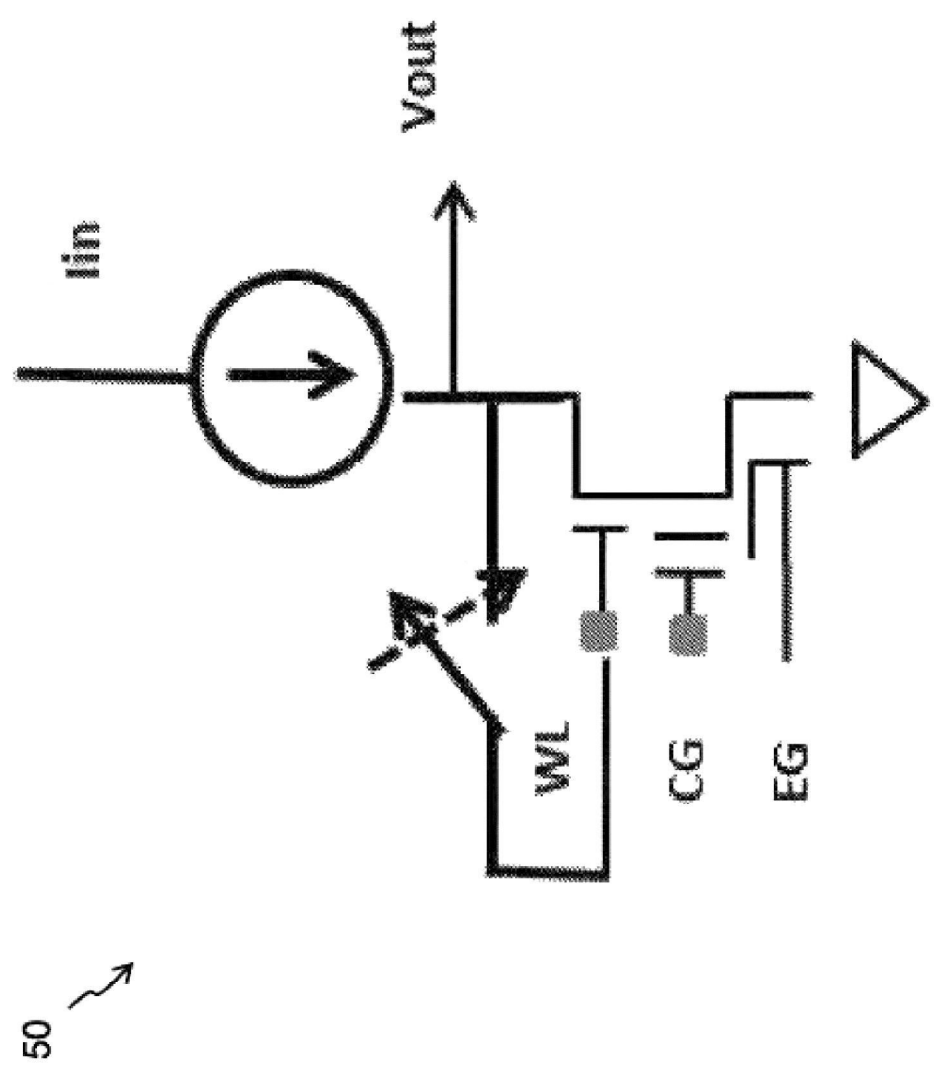
【圖22】



【圖23】

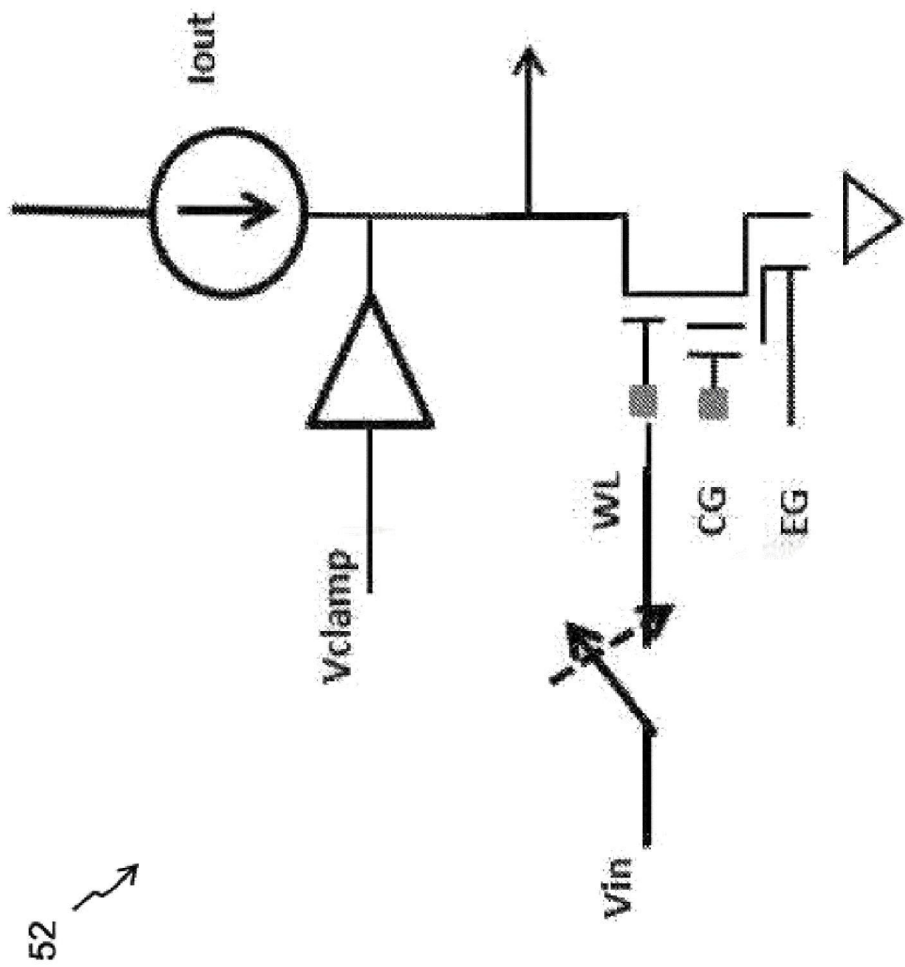


【圖24】



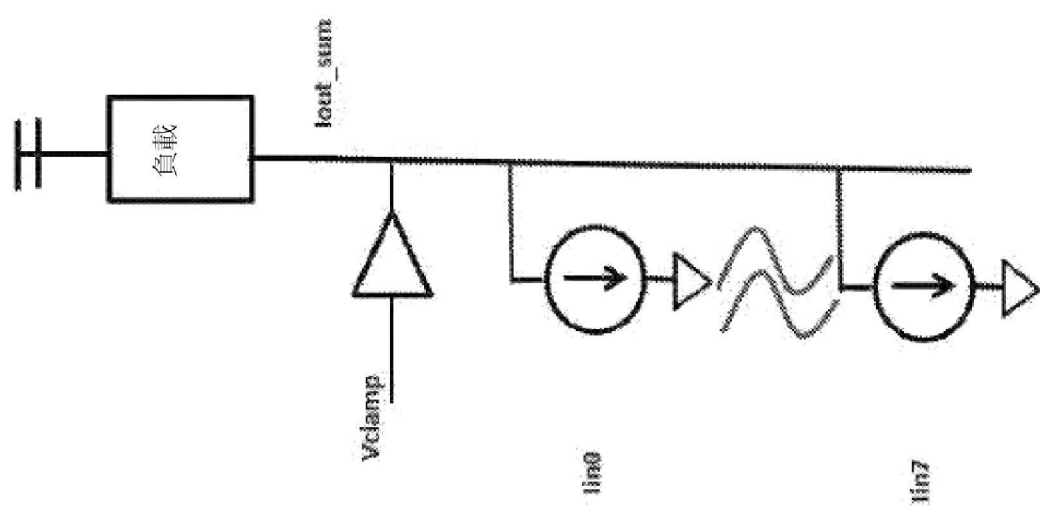
【圖25】

50 ↗



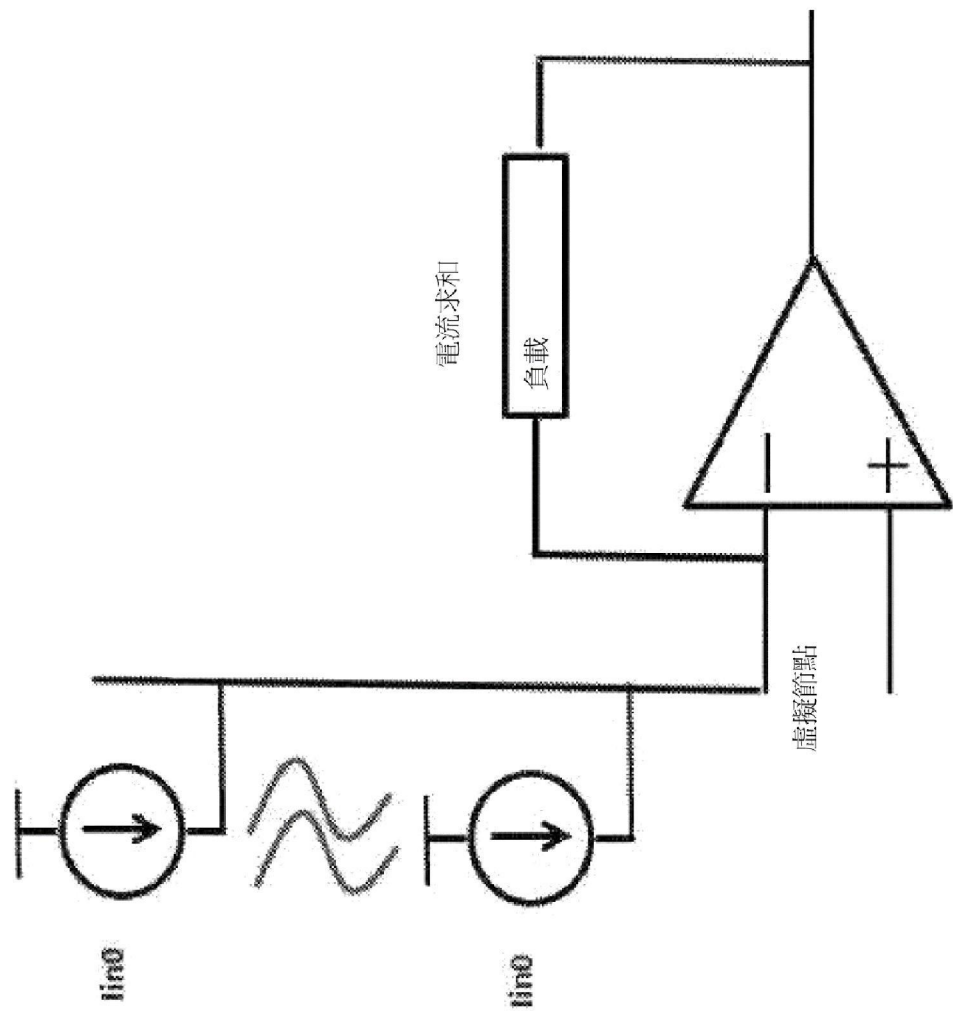
【圖26】

52 ↗



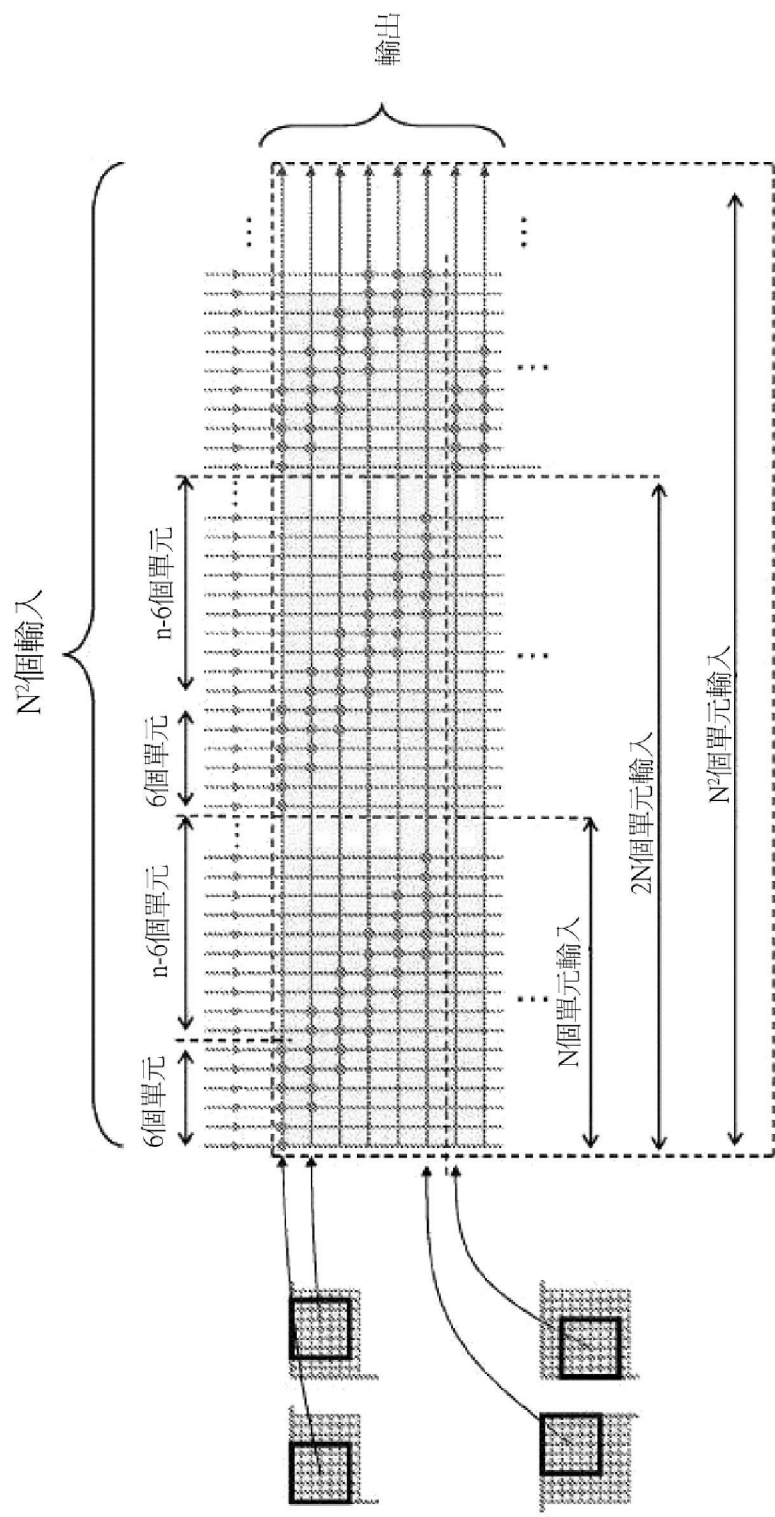
【圖27】

54 ↗

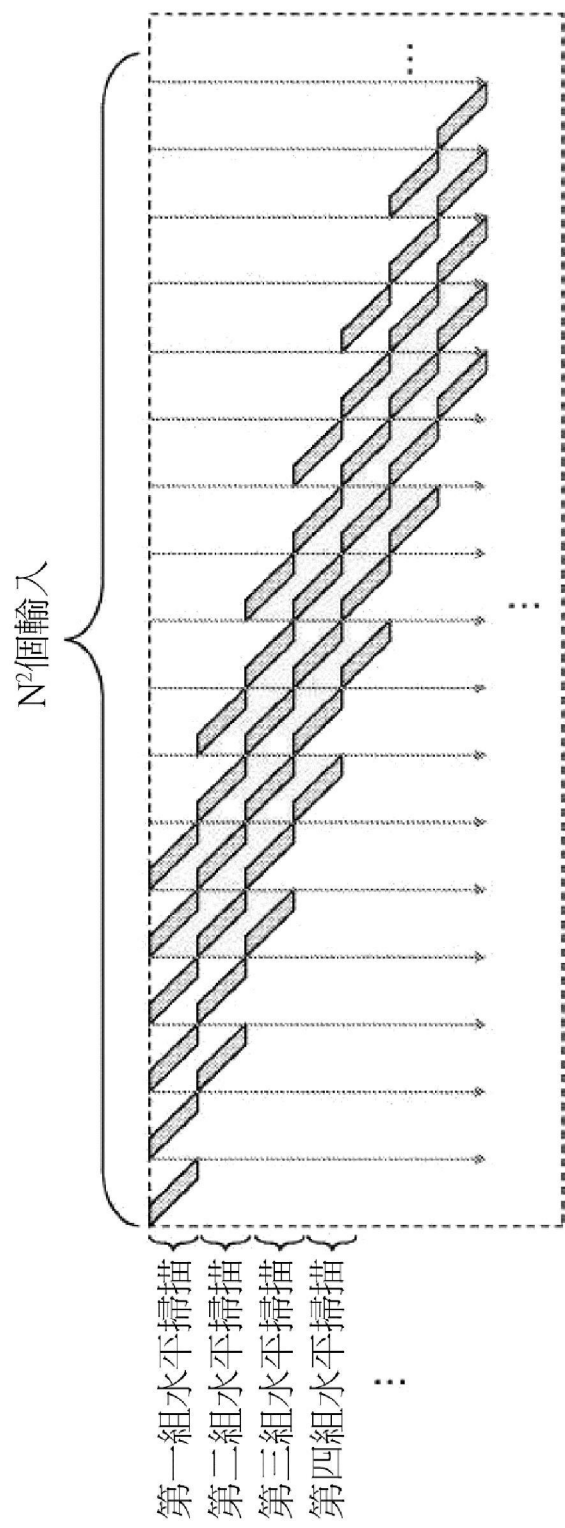


【圖28】

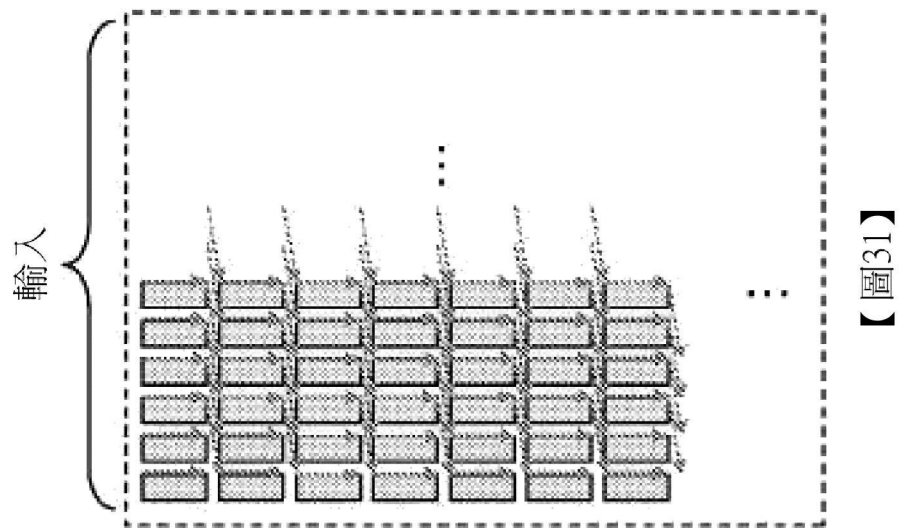
56 ↗

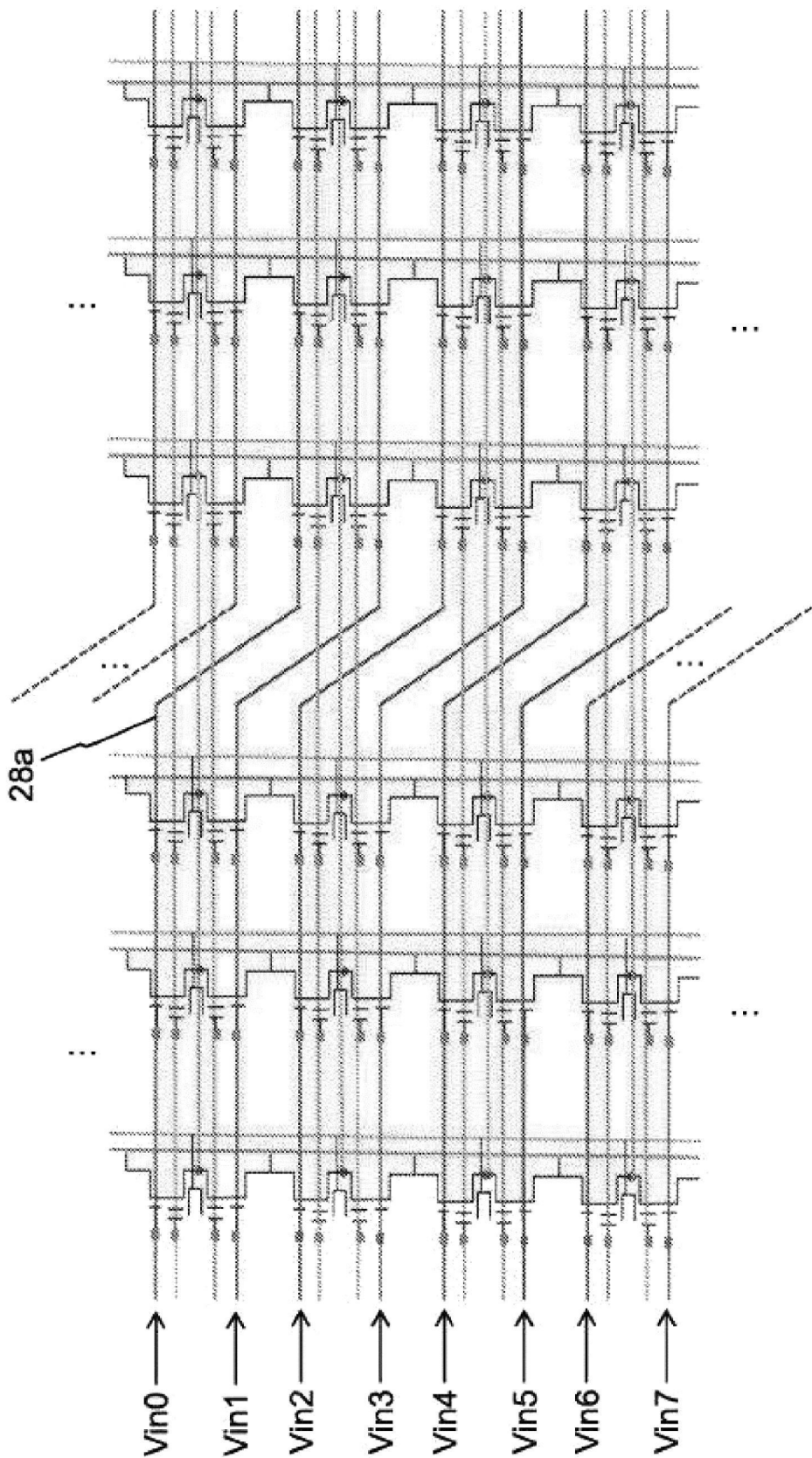


【圖29】

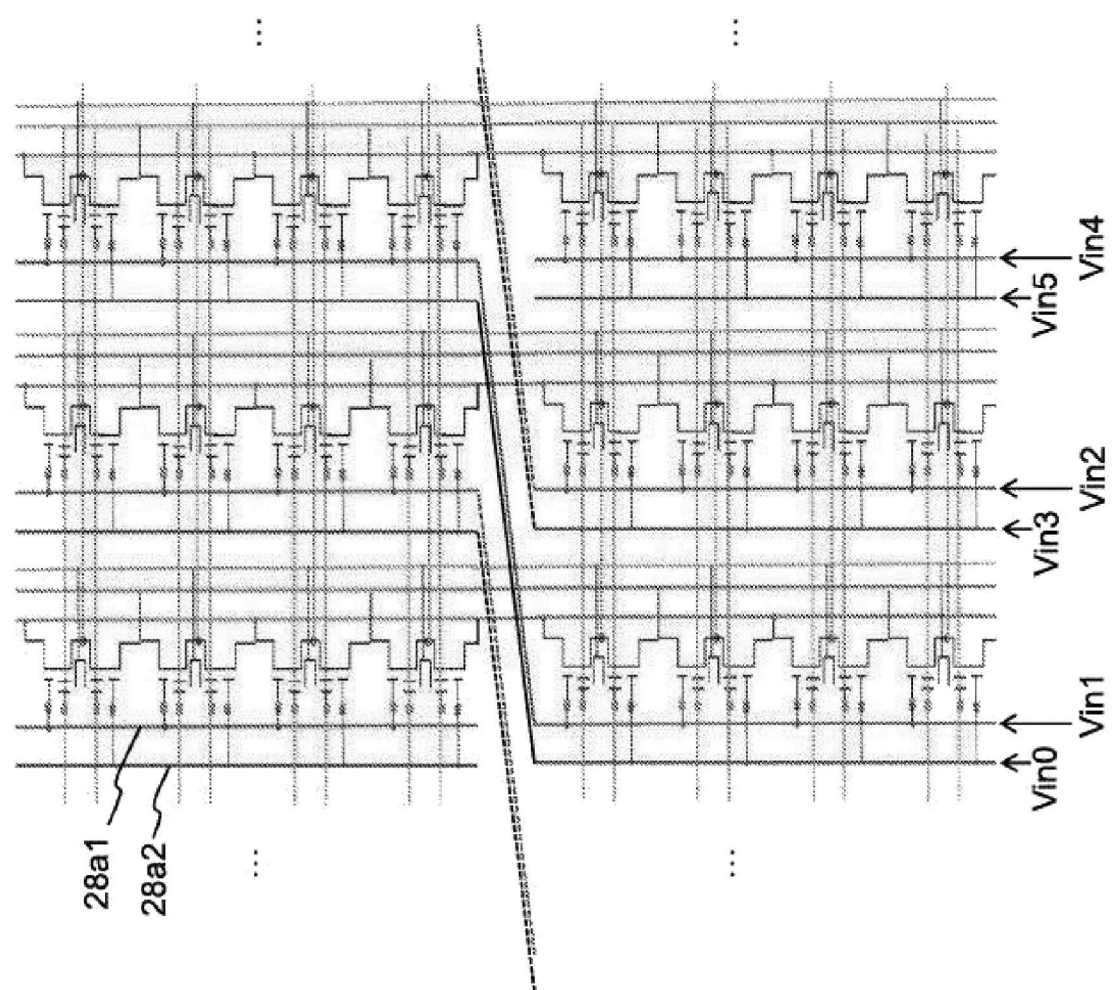


【圖30】

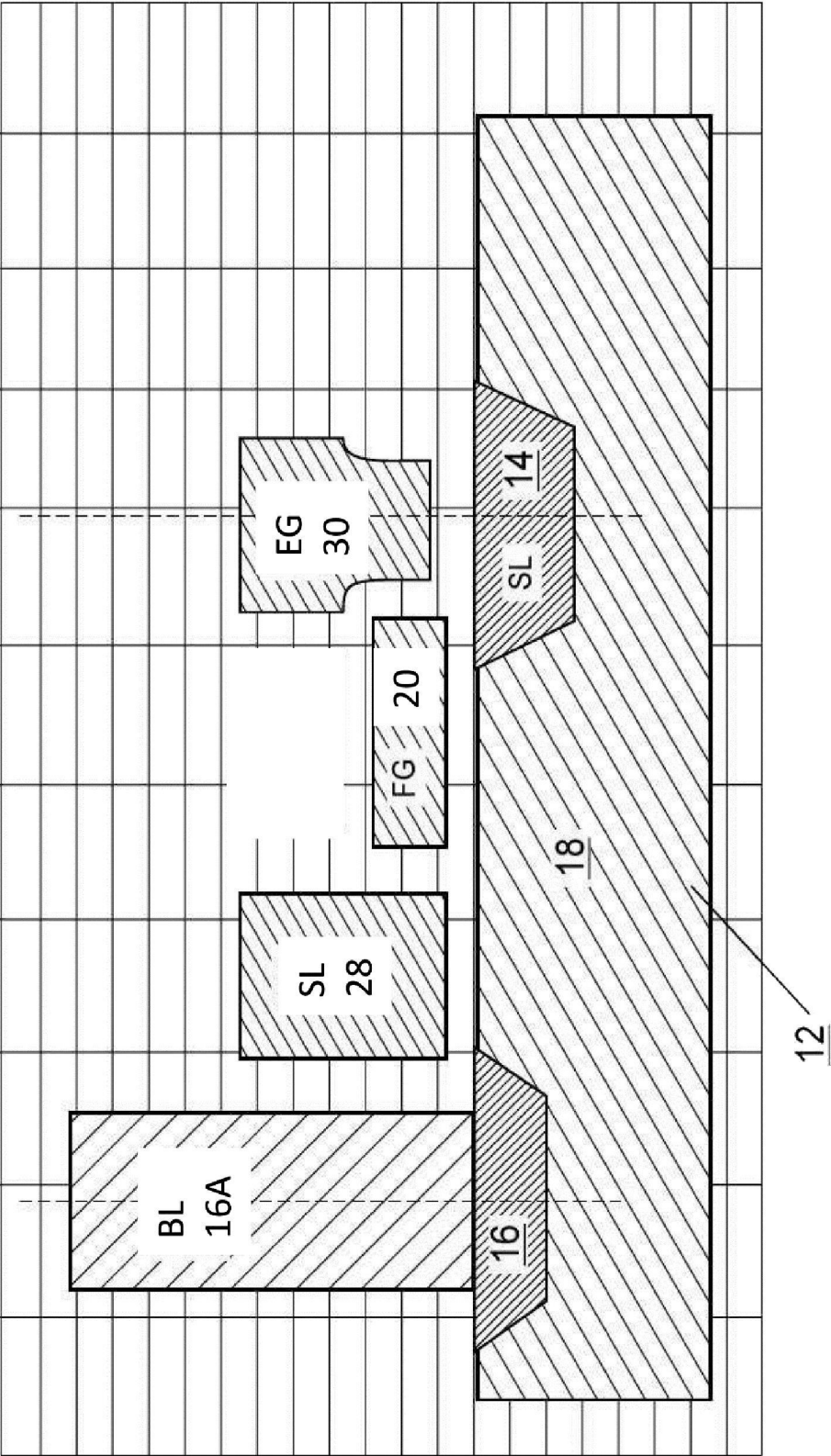




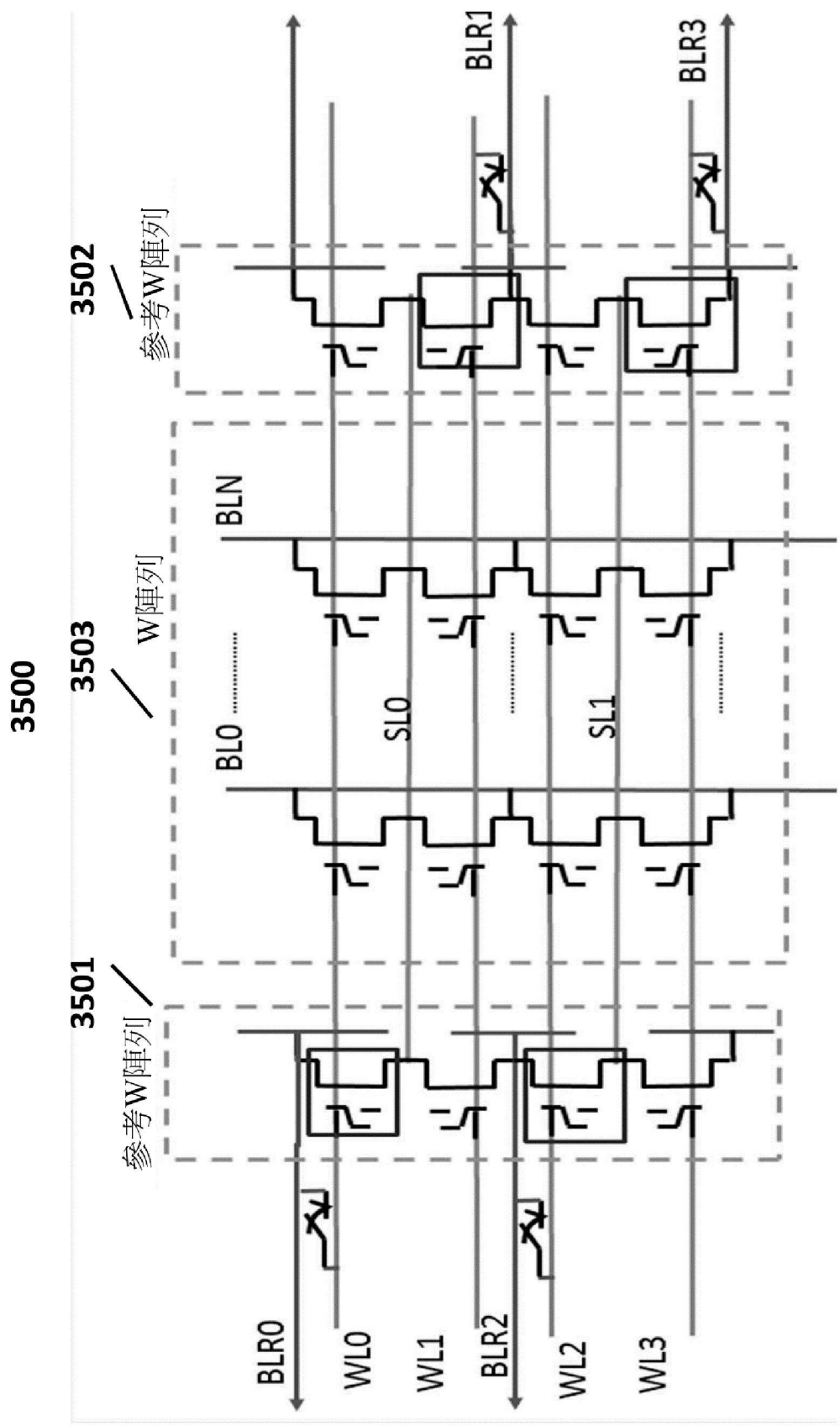
【圖32】



【圖33】



【圖34（先前技術）】

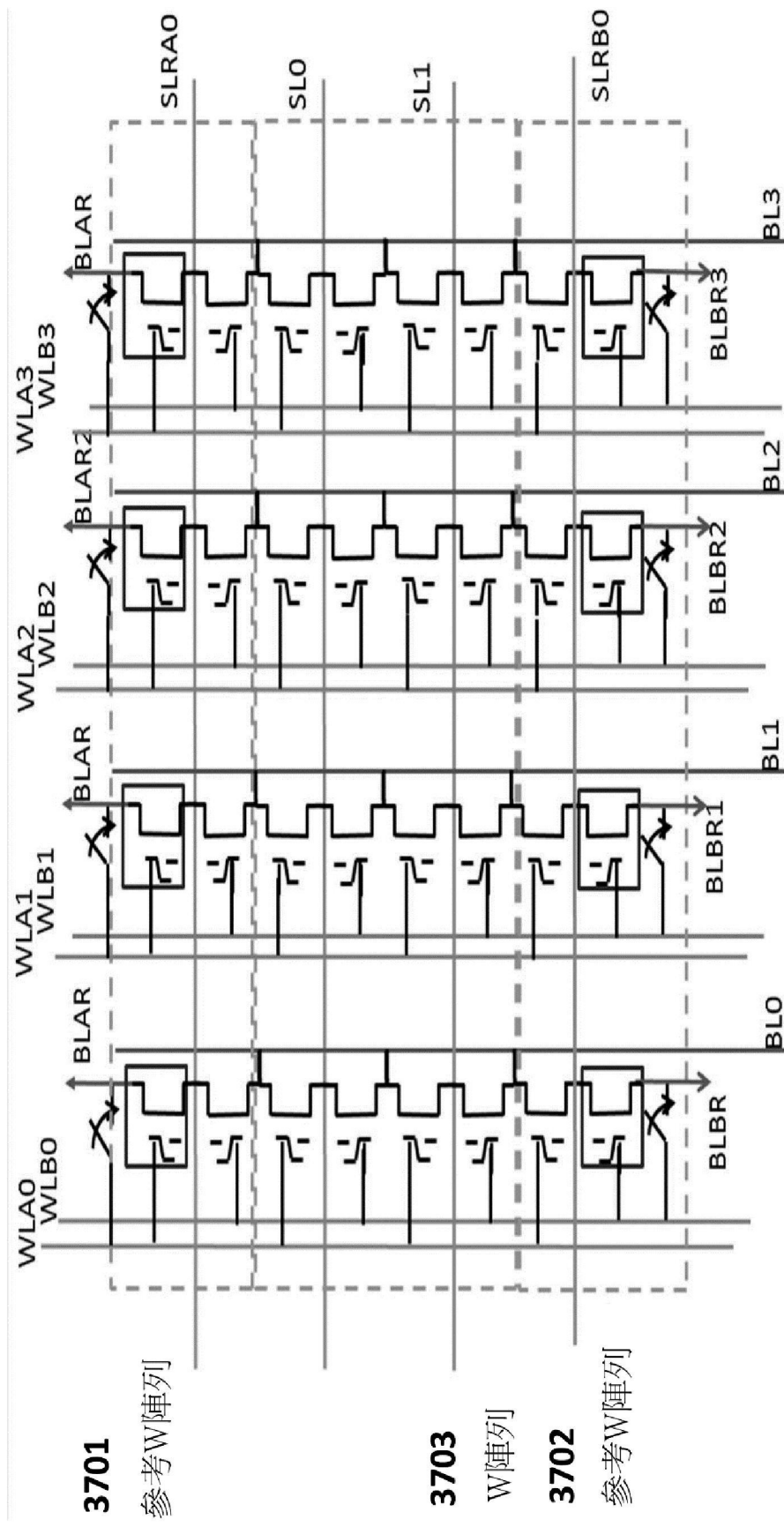


【圖35】

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

【圖36】

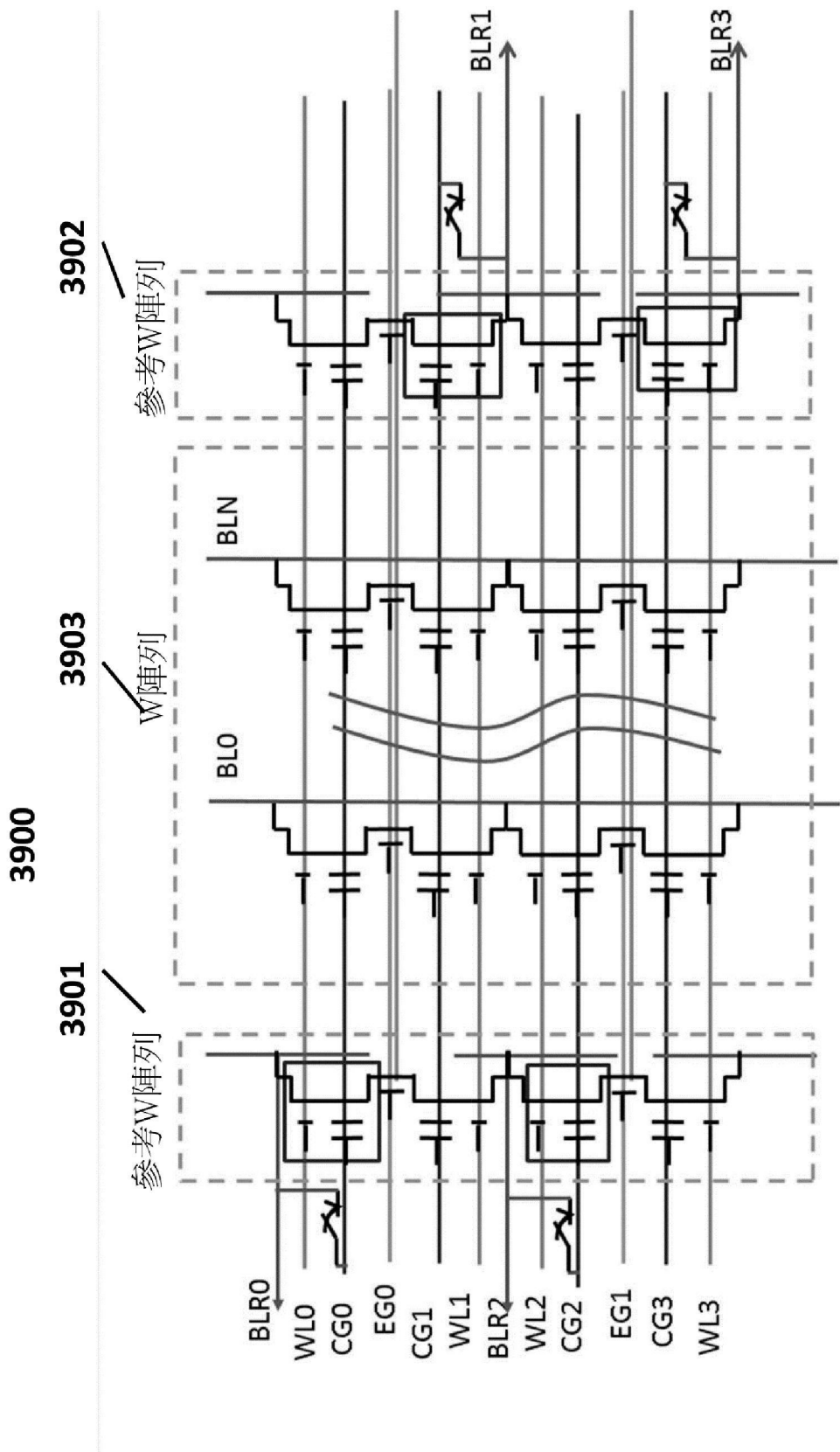
3700



【圖37】

	WL	WL-非選取	BL	BL-非選取	SL	SL-非選取
讀取	1至3.5 V	-0.5 V/0V	0.6至2 V	0.6 V至2 V/0 V	~1 V (Ineuron)	0 V
抹除	~5至13 V	0 V	0 V	0 V	0 V	SL-inhibit (~4至8 V)
程式化	1至2 V	-0.5 V/0V	0.1至3 uA	Vinh ~2.5 V	4至10 V	0至1 V/FLT

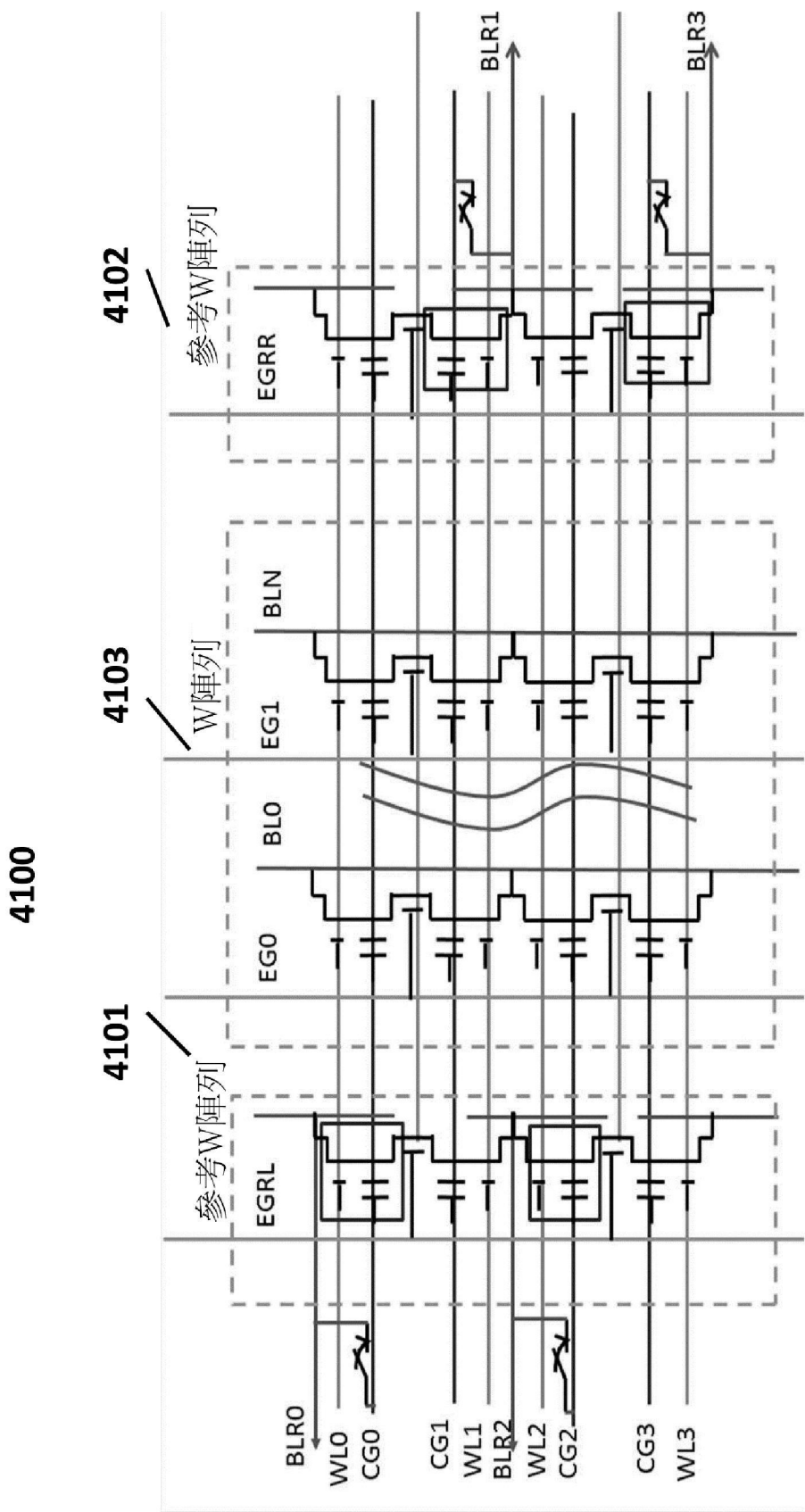
【圖38】



【圖39】

	WL	WL-非選取	BL	BL-非選取	CG	CG-非選取 相同區段	CG-非選取	EG	EG-非選取	SL	SL-非選取
讀取	1.0至2 V	-0.5 V/ 0V	0.6至2 V (lneuron)	0 V	0至2.6 V	0至2.6 V	0至2.6 V	0至2.6 V	0至2.6 V	0 V	0 V
抹除	0 V	0 V	0 V	0 V	0 V	0至2.6 V	0至2.6 V	5至12 V	0至2.6 V	0 V	0 V
程式化	0.7至1 V	-0.5 V/ 0V	0.1至1 uA	Vinh (1至2 V)	4-11 V	0至2.6 V	0至2.6 V	4.5至5 V	0至2.6 V	4.5至5 V	0至1 V

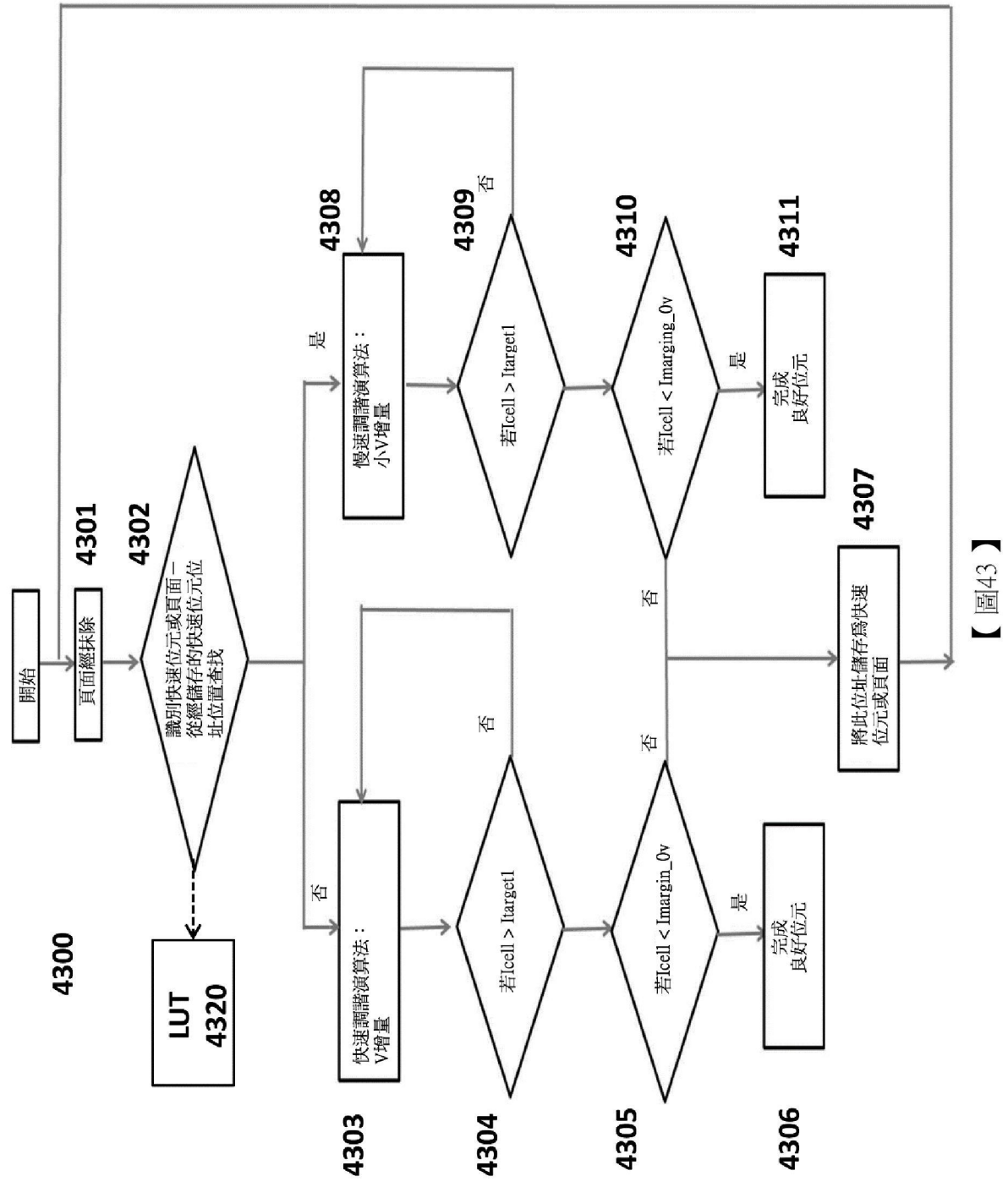
【圖40】

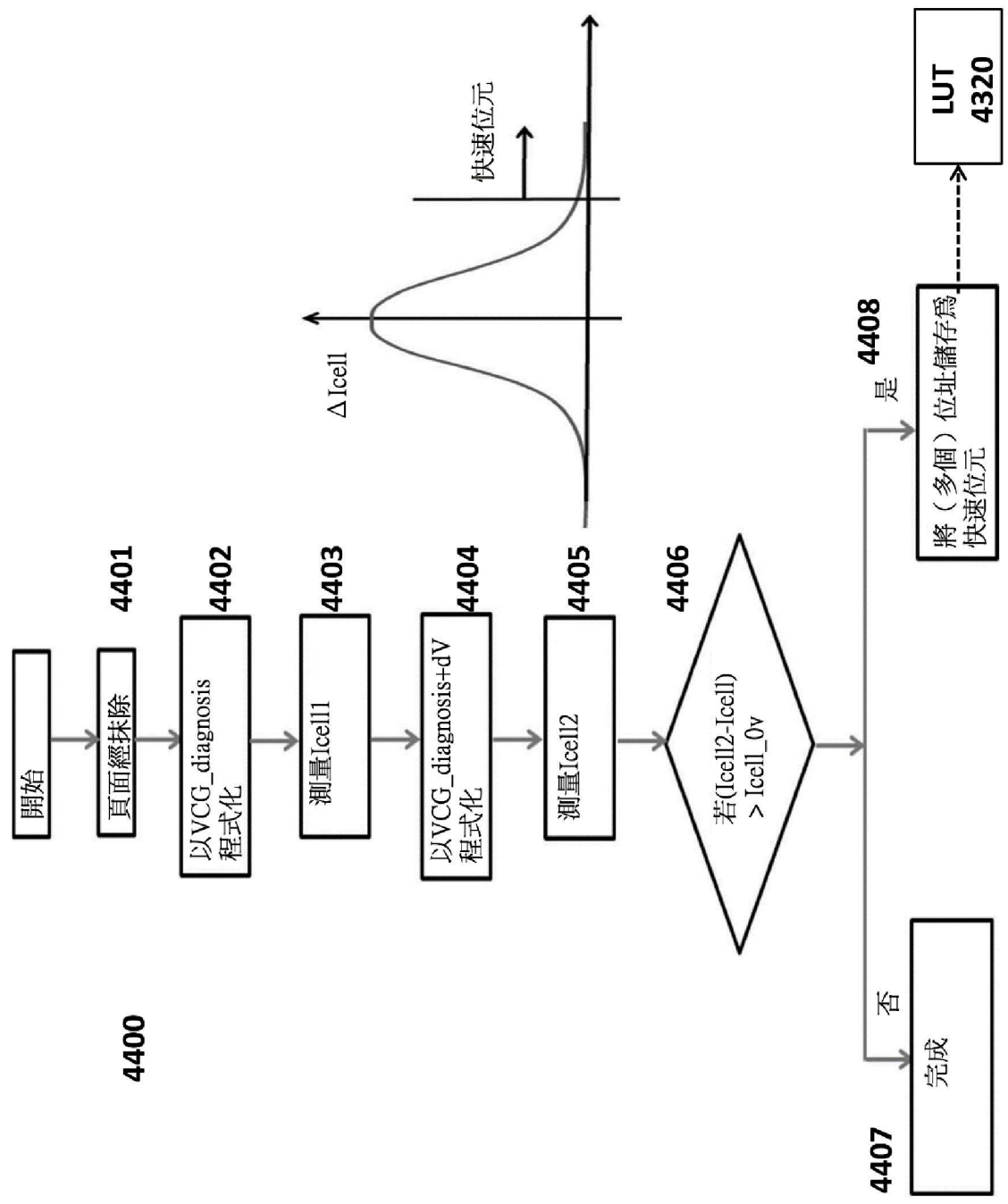


【圖41】

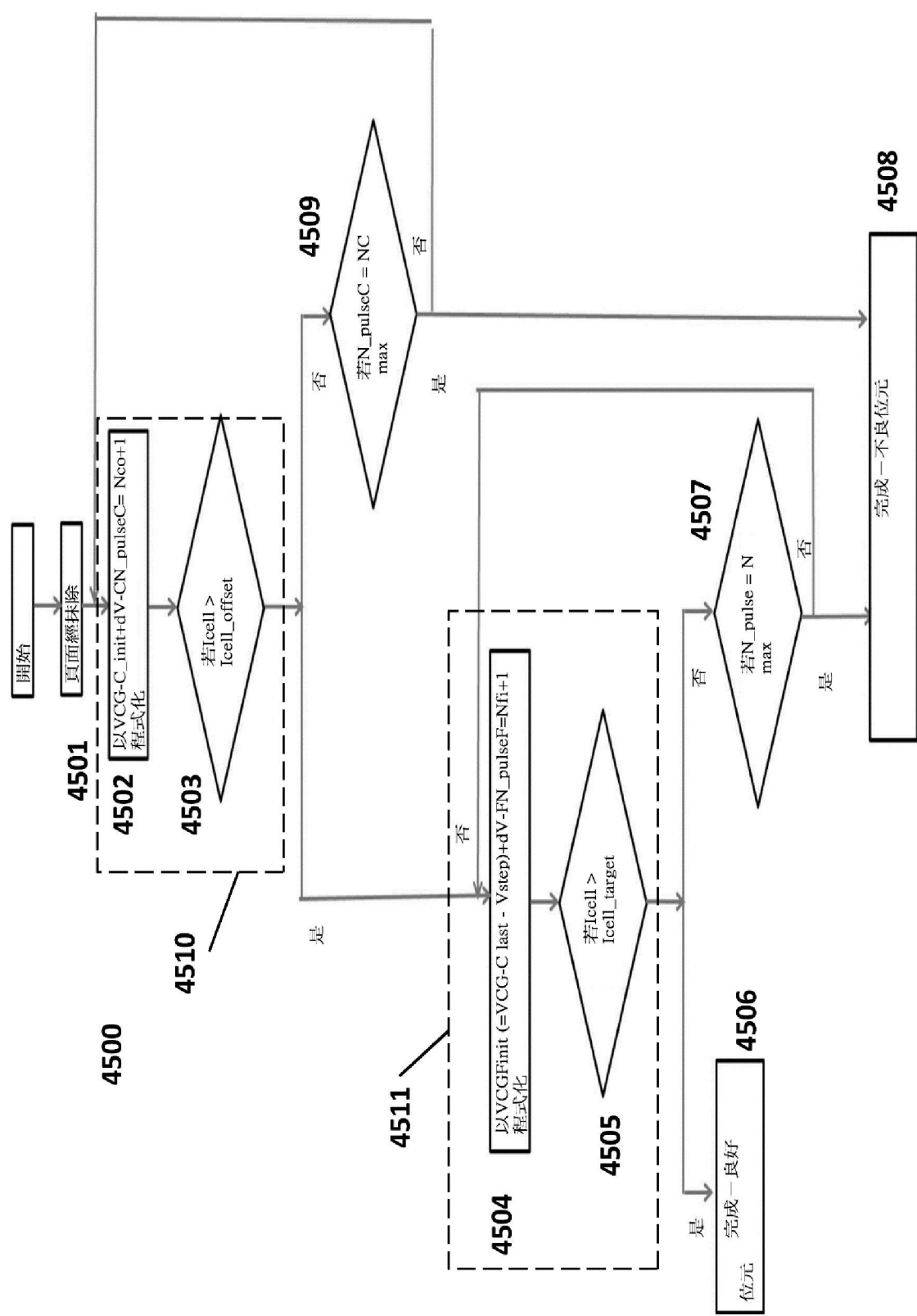
	WL	WL-非選取	BL	BL-非選取	CG	CG-非選取 相同區段	CG-非選取	EG	EG-非選取	SL	SL-非選取
讀取	1.0至2 V	-0.5 V/ 0V	0.6至2 V (Inuron)	0 V	0至2.6 V	0至2.6 V	0至2.6 V	0至2.6 V	0至2.6 V	0 V	0 V
抹除	0 V	0 V	0 V	0 V	0 V	CGINH(4至9 V)	0至2.6 V	5至12 V	0至2.6 V	0 V	0 V
程式化	0.7至1 V	-0.5 V/ 0V	0.1至1 uA	Vinh (1至2 V)	4-11 V	0至2.6 V	0至2.6 V	4.5至5 V	0至2.6 V	4.5至5 V	0至1 V

【圖42】





【圖44】



【圖45】

均勻步階演算法
4600

V target		8 V
V start		4 V
均勻步階		
	步階大小	0.01 V
	步階數	400
步階總數：		400

【圖46】

均勻對數步階
粗略/精細演算法

4700

V target		8 V
V start		4 V
粗略－均勻步階		
	步階大小	0.4 V
	步階數	10
精組1		
	Istepdown	0.2 V
	步階大小	0.1 V
	步階數	6
精組2		
	Istepdown	0.1 V
	步階大小	0.01 V
	步階數	20
步階總數：		46

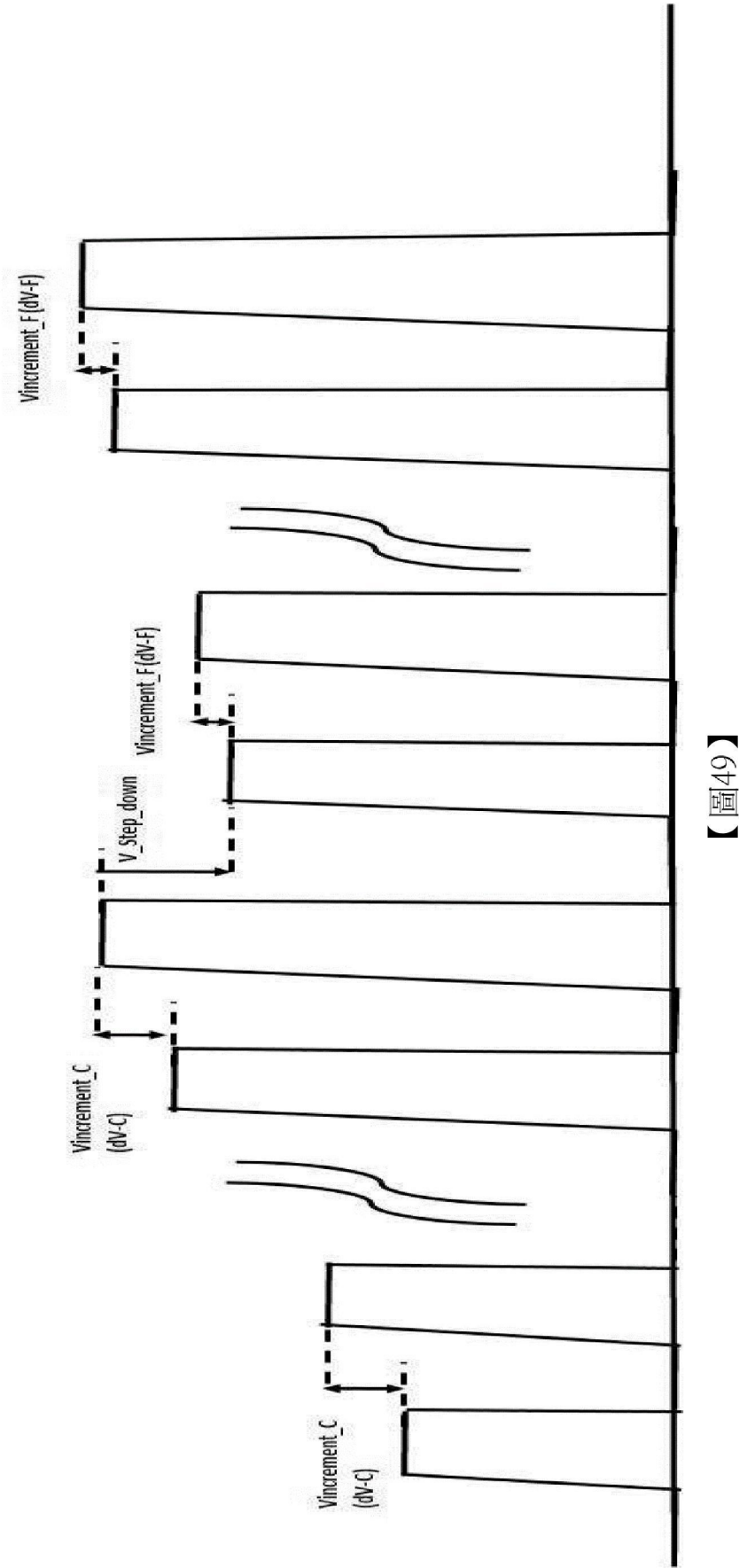
【圖47】

二元搜尋步階
粗略/精細演算法

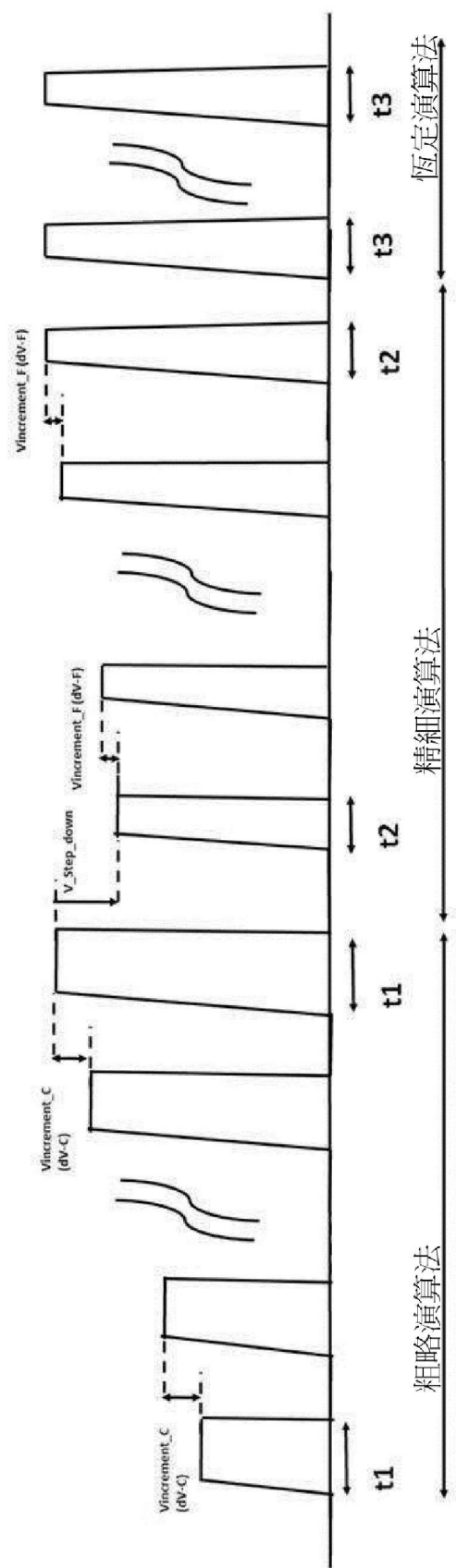
4800

V target		8 V
V start		4 V
粗略－二元搜尋步階		
	step1	2 V
	step2	1 V
	step3	0.5 V
	step4	0.25 V
	step5	0.125 V
	step6	0.0625 V
精細1	Istepdown	0.0625 V
	步階大小	0.01 V
	步階數	19
		25

【圖48】



【圖49】



【圖50】