



(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 I821746 B

(45)公告日：中華民國 112 (2023) 年 11 月 11 日

(21)申請案號：110134382

(22)申請日：中華民國 110 (2021) 年 09 月 15 日

(51)Int. Cl.：

G11C11/40 (2006.01)**G11C11/54 (2006.01)****G06F17/16 (2006.01)****G06N3/06 (2006.01)****G06T1/40 (2006.01)**

(30)優先權：2021/02/02

美國

63/144,477

(71)申請人：旺宏電子股份有限公司 (中華民國) MACRONIX INTERNATIONAL CO., LTD.

(TW)

新竹縣科學工業園區力行路 16 號

(72)發明人：呂函庭 LUE, HANG-TING (TW)；許柏凱 HSU, PO-KAI (TW)

(74)代理人：祁明輝；林素華

(56)參考文獻：

TW 201941209A

US 2019/0237142A1

審查人員：林育弘

申請專利範圍項數：8 項 圖式數：4 共 20 頁

(54)名稱

記憶體裝置及其操作方法

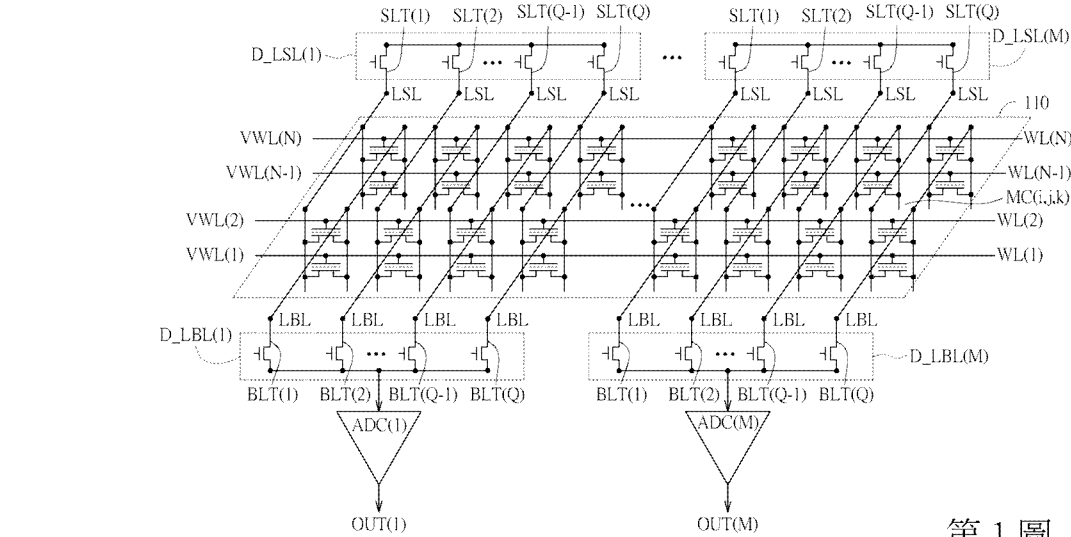
(57)摘要

本案提出記憶體裝置及其操作方法。該操作方法包括：於進行一乘積累加運算操作時，透過複數個第一信號線輸入複數個輸入至該記憶體裝置之複數個記憶體單元；根據該些記憶體單元的複數個權重，該些記憶體單元輸出複數個單元電流於複數個第二信號線；加總各該些第二信號線上的該些單元電流成複數個信號線電流；加總該些信號線電流成一整體信號線電流；以及將該整體信號線電流轉換成一輸出，其中，該輸出代表該些輸入與該些權重的一乘積累加運算操作結果。

A memory device and an operation method thereof are provided. The operation method comprises: in performing a multiply accumulate (MAC) operation, inputting a plurality of inputs into a plurality of memory cells via a plurality of first signal lines; outputting a plurality of cell currents from the memory cells to a plurality of second signal lines based on a plurality of weights of the memory cells; summing the cell currents on each of the second signal lines into a plurality of signal line currents; summing the signal line currents into a global signal line current; and converting the global signal line current into an output, wherein the output represents a MAC operation result of the inputs and the weights.

指定代表圖：

100



第 1 圖

符號簡單說明：

100:記憶體裝置

110:記憶體陣列

D_LBL(1)~D_LBL(M):區域位元線解碼器

D_LSL(1)~D_LSL(M):區域源極線解碼器

ADC(1)~ADC(M):轉換單元

BLT(1)~BLT(Q):位元線電晶體

SLT(1)~SLT(Q):源極線電晶體

MC(i,j,k):記憶體單元

WL(1)~WL(N):字元線

LBL:區域位元線

LSL:區域源極線



I821746

【發明摘要】

【中文發明名稱】 記憶體裝置及其操作方法

【英文發明名稱】 MEMORY DEVICE AND OPERATION

METHOD THEREOF

【中文】

本案提出記憶體裝置及其操作方法。該操作方法包括：於進行一乘積累加運算操作時，透過複數個第一信號線輸入複數個輸入至該記憶體裝置之複數個記憶體單元；根據該些記憶體單元的複數個權重，該些記憶體單元輸出複數個單元電流於複數個第二信號線；加總各該些第二信號線上的該些單元電流成複數個信號線電流；加總該些信號線電流成一整體信號線電流；以及將該整體信號線電流轉換成一輸出，其中，該輸出代表該些輸入與該些權重的一乘積累加運算操作結果。

【英文】

A memory device and an operation method thereof are provided. The operation method comprises: in performing a multiply accumulate (MAC) operation, inputting a plurality of inputs into a plurality of memory cells via a plurality of first signal lines; outputting a plurality of cell currents from the memory cells to a plurality of second signal lines based on a plurality of weights of the memory cells; summing the cell currents on each of the second signal lines into a plurality of signal line currents; summing the signal line currents into a global signal line current; and converting the global signal line current into an

output, wherein the output represents a MAC operation result of the inputs and the weights.

【指定代表圖】 第 1 圖。

【代表圖之符號簡單說明】

100:記憶體裝置

110: 記憶體陣列

D_LBL(1)~ D_LBL(M): 區域位元線解碼器

D_LSL(1)~ D_LSL(M): 區域源極線解碼器

ADC(1)~ADC(M): 轉換單元

BLT(1)~BLT(Q): 位元線電晶體

SLT(1)~SLT(Q): 源極線電晶體

MC(i, j, k): 記憶體單元

WL(1)~WL(N):字元線

LBL: 區域位元線

LSL: 區域源極線

【特徵化學式】

無

【發明說明書】

【中文發明名稱】 記憶體裝置及其操作方法

【英文發明名稱】 MEMORY DEVICE AND OPERATION

METHOD THEREOF

【技術領域】

【0001】 本發明是有關於一種記憶體裝置及其操作方法。

【先前技術】

【0002】 人工智慧(AI)日漸重要。乘積累加運算(Multiply Accumulate, MAC)操作是AI的核心操作。

【0003】 在傳統上，為完成MAC操作，要透過算術邏輯單元(Arithmetic logic unit, ALU)、浮點運算器等，把資料從記憶體存取出來進行運算，這需要大量資料搬移，故而，運算速度較慢。

【0004】 現已發展出記憶體內運算(Computing-in-Memory, CIM)記憶體，以求快速完成MAC，適合用於實施AI加速器。

【0005】 以目前而言，記憶體裝置已朝向3D堆疊發展，以提高記憶體密度。以3D結構而言，除了3D NAND快閃記憶體與3D NOR快閃記憶體之外，目前又已發展出3D AND快閃記憶體。

【0006】 如何在3D記憶體中，在不額外佔電路面積的前提下，提高MAC運算量，乃是業界努力方向之一。

【發明內容】

【0007】 根據本案一實施例，提出一種記憶體裝置之操作方法，該操作方法包括：於進行一乘積累加運算(Multiply Accumulate, MAC)操作時，透過複數個第一信號線輸入複數個輸入至該記憶體裝置之複數個記憶體單元；根據該些記憶體單元的複數個權重，該些記憶體單元輸出複數個單元電流於複數個第二信號線；加總各該些第二信號線上的該些單元電流成複數個信號線電流；加總該些信號線電流成一整體信號線電流；以及將該整體信號線電流轉換成一輸出，其中，該輸出代表該些輸入與該些權重的一乘積累加運算操作結果。

【0008】 根據本案又一實施例，提出一種記憶體裝置，包括：一記憶體陣列，包括複數個記憶體單元，該些記憶體單元儲存複數個權重，該些記憶體單元耦接至複數個第一信號線與複數個第二信號線；至少一第一區域信號線解碼器，耦接至該記憶體陣列與至少一第一整體信號線；以及至少一轉換單元，耦接至該至少一第一區域信號線解碼器與該至少一第一整體信號線。

【0009】 為了對本發明之上述及其他方面有更佳的瞭解，下文特舉實施例，並配合所附圖式詳細說明如下：

【圖式簡單說明】**【0010】**

第1圖繪示根據本案一實施例的記憶體裝置的電路示意圖。

第2圖顯示根據本案一實施例的記憶體裝置進行MAC操作時

的示意圖。

第3圖顯示根據本案一實施例的記憶體操作方法流程圖。

第4A圖至第4D圖顯示根據本案一實施例的裝置性能特徵圖。

【實施方式】

【0011】本說明書的技術用語係參照本技術領域之習慣用語，如本說明書對部分用語有加以說明或定義，該部分用語之解釋係以本說明書之說明或定義為準。本揭露之各個實施例分別具有一或多個技術特徵。在可能實施的前提下，本技術領域具有通常知識者可選擇性地實施任一實施例中部分或全部的技術特徵，或者選擇性地將這些實施例中部分或全部的技術特徵加以組合。

【0012】請參照第 1 圖，其繪示根據本案一實施例的記憶體裝置的電路示意圖。如第 1 圖所示，本案一實施例的記憶體裝置 100 包括：記憶體陣列 110，複數個區域位元線解碼器 $D\_LBL(1) \sim D\_LBL(M)$ (M 為正整數)、複數個區域源極線解碼器 $D\_LSL(1) \sim D\_LSL(M)$ ，與複數個轉換單元 $ADC(1) \sim ADC(M)$ 。在此，轉換單元為類比數位轉換單元為例做說明，但當知本案並不受限於此。記憶體裝置 100 例如但不受限於，為 3D(三維)AND 型記憶體裝置，而記憶體陣列 110 為 3D AND 型記憶體陣列。

【0013】各區域位元線解碼器 $D\_LBL(1) \sim D\_LBL(M)$ 包括複數個位元線電晶體 $BLT(1) \sim BLT(Q)$ (Q 為正整數)。相似地，各區域源極線解碼器 $D\_LSL(1) \sim D\_LSL(M)$ 包括複數個源極線電

晶體 $SLT(1) \sim SLT(Q)$ 。

【0014】記憶體陣列 110 包括以陣列排列的複數個記憶體單元 $MC(i, j, k)$ 。該些記憶體單元 $MC(i, j, k)$ 耦接至複數條字元線 $WL(1) \sim WL(N)$ (N 為正整數)、複數條區域源極線 LSL 與複數條區域位元線 LBL 。 $i=1 \sim N$ ， $j=1 \sim M$ ， $k=1 \sim Q$ 。 i 、 j 與 k 為正整數。

【0015】以位元線電晶體 $BLT(1)$ 為例，位元線電晶體 $BLT(1)$ 具有：一第一端(如源極)耦接至區域位元線 LBL ，一第二端(如汲極)耦接至轉換單元之輸入端與一整體位元線(未示出)，以及一控制端(如閘極)接收一控制信號(未示出)。位元線電晶體 $BLT(2) \sim BLT(Q)$ 具有類似耦接關係。

【0016】相似地，以源極線電晶體 $SLT(1)$ 為例，源極線電晶體 $SLT(1)$ 具有：一第一端(如源極)耦接至區域源極線 LSL ，一第二端(如汲極)耦接至一整體源極線(未示出)，以及一控制端(如閘極)接收一控制信號(未示出)。源極線電晶體 $SLT(2) \sim SLT(Q)$ 具有類似耦接關係。

【0017】當進行乘積累加運算(Multiply Accumulate, MAC)時，字元線 $WL(1) \sim WL(N)$ 接收字元線電壓 $VWL(1) \sim VWL(N)$ ，其中，字元線電壓 $VWL(1) \sim VWL(N)$ 為高位準電壓或低位準電壓。當進行 MAC 運算時，該些字元線電壓 $VWL(1) \sim VWL(N)$ 即為輸入。

【0018】該些記憶體單元可被程式化為邏輯 1 或邏輯 0，亦

即，在本案一實施例中，該些記憶體單元乃是單階儲存單元 (Single-Level Cell, SLC)，但本案並不限於此。在本案其他可能實施例中，該些記憶體單元可為多階儲存單元 (Multi-Level Cell, MLC)，此亦在本案精神範圍內。當該記憶體單元被程式化為邏輯 1 時且相關的字元線上被施加高位準電壓時，則該記憶體單元會輸出單元電流 (cell current)；當該記憶體單元被程式化為邏輯 1 時且相關的字元線上被施加低位準電壓時，則該記憶體單元不會輸出單元電流；以及，當該記憶體單元被程式化為邏輯 0 時，不論相關的字元線上被施加高位準電壓或低位準電壓時，該記憶體單元不會輸出單元電流。該記憶體單元 $MC(i, j, k)$ 所輸出的單元電流 $I_{cell}(i, j, k)$ 可表示為 $I_{cell}(i, j, k) = VWL(i) * w(i, j, k)$ ，其中， $w(i, j, k)$ 代表該記憶體單元 $MC(i, j, k)$ 所儲存的權重值，亦即，該記憶體單元 $MC(i, j, k)$ 的跨導值 (transconductance)。

【0019】故而，以同一條區域位元線 LBL 而言，從該條區域位元線 LBL 流向該轉換單元 $ADC(j)$ 的位元線電流 (信號線電流) 乃是在該條區域位元線 LBL 上的 N 個記憶體單元的單元電流的總和。

【0020】各區域位元線解碼器 $D\_LBL(1) \sim D\_LBL(M)$ 將該些區域位元線 LBL 上的位元線電流 (信號線電流) 加總成一整體位元線電流 (亦可稱為整體信號線電流)。故而，由此可推出，整體位元線電流 = $\sum_{i=1}^N \sum_{k=1}^{k=Q} VWL(i) * w(i, j, k)$ 。

第 5 頁，共 10 頁 (發明說明書)

【0021】轉換單元 ADC(1)~ADC(M)接收區域位元線解碼器 D_LBL(1)~ D_LBL(M)的個別整體位元線電流並轉換成輸出(數位碼)後得到輸出 OUT(1)~OUT(M)。例如但不受限於，當轉換單元 ADC(1)~ADC(M)具有 8 位元解析度，則可將輸入電流轉換成 8 位元的輸出 OUT(1)~OUT(M)。故而，輸出 OUT(j)可表示為： $\sum_{i=1}^N \sum_{k=1}^Q IN(i) * w(i,j,k)$ ，其中，IN(i)代表輸入至記憶體陣列 110 的字元線 WL(i)的輸入資料。當輸入資料 IN(i)為邏輯高時，字元線電壓 VWL(i)為高準位電壓；以及當輸入資料 IN(i)為邏輯低時，字元線電壓 VWL(i)為低準位電壓。

【0022】亦即，轉換單元 ADC(j)的輸出 OUT(j)相關於耦接至同一轉換單元 ADC(j)的該些記憶體單元的儲存權重與該些相關字元線電壓(輸入資料)的 MAC 操作結果。

【0023】現請參照第 2 圖，其顯示根據本案一實施例的記憶體裝置進行 MAC 操作時的示意圖。如第 2 圖所示，於進行 MAC 操作時，該些位元線電晶體 BLT(1)~BLT(3)與該些源極線電晶體 SLT(1)~SLT(3)為導通，且施加至整體位元線 GBLj 的整體位元線電壓為 1.8V，而施加至整體源極線 GSLj 的整體源極線電壓為 0V。字元線電壓 VWL(1)~VWL(4)的高位準電壓為 2.8V，而低位準電壓為 0V。

【0024】故而，在第 2 圖中，整體位元線電流 = $\sum_{i=1}^4 \sum_{k=1}^3 VWL(i) * w(i,j,k)$ 。

【0025】由第 2 圖可看出，由該 3 個位元線電晶體

BLT(1)~BLT(3)所加總的電流可代表多階權重(multi-level weight)0、1、2 與 3，亦即，0 代表 2 階的 00，1 代表 2 階的 01，2 代表 2 階的 10 與 3 代表 2 階的 11。各記憶體單元乃儲存單階權重 1 或 0。

【0026】進一步說，當想要代表 x 階權重時，則耦接至同一轉換單元的區域位元線數量為： $Q=2^x-1$ 。例如，想要代表 4 階權重時，則耦接至同一轉換單元的區域位元線數量為： $Q=2^4-1=15$ 。

【0027】也就是說，在本案一實施例中，即便是使用單階儲存單元，仍可進行多階權重的運算。故而，本案實施例具有架構簡單但可執行複雜 MAC 運算的優點。

【0028】第 3 圖顯示根據本案一實施例的記憶體操作方法流程圖。於步驟 310 中，於進行一乘積累加運算(Multiply Accumulate, MAC)操作時，透過複數個第一信號線輸入複數個輸入至該記憶體裝置之複數個記憶體單元。於步驟 320 中，根據該些記憶體單元的複數個權重，該些記憶體單元輸出複數個單元電流於複數個第二信號線。於步驟 330 中，加總各該些第二信號線上的該些單元電流成複數個信號線電流。於步驟 340 中，加總該些信號線電流成一整體信號線電流。於步驟 350 中，將該整體信號線電流轉換成一輸出，其中，該輸出代表該些輸入與該些權重的一乘積累加運算操作結果。

【0029】第 4A 圖至第 4D 圖顯示根據本案一實施例的裝置性能特徵圖。如第 4A 圖所示，在本案一實施例中，如果可以讓記

憶體單元的導通電流(I_{on})與關閉電流(I_{off})之間的差異愈大的話(例如, $(I_{on}/I_{off}) > 10^4$), 則可以允許更多的平行加總(parallel summation)與降低背景漏電流(background leakage)。此外, 透過步增程式脈衝(ISPP, Increment Step Programming Pulse)可以逐漸將臨界電壓抬高。當字元線電壓固定為 2.8V 時, 可逐漸地將單元電流修改地更小。

【0030】第 4B 圖顯示可調整且緊縮(tunable and tight)的單元電流。如第 4B 圖所示, 可將單元電流 I_{cell} 修改成不同範圍, 例如但不受限於, 從單元電流 I_{cell} 的範圍可從 150nA 至 1.5 μA 。而且, 單元電流 I_{cell} 的分布更加緊縮, 單元電流 I_{cell} 的標準差(Standard Deviation, 數學符號 σ (sigma))可小於 2% ($\sigma < 2\%$)。

【0031】第 4C 圖顯示根據本案一實施例中, 3D AND 型記憶體裝置可免於讀取干擾(Read-disturb free), 例如是當字元線電壓約為 +7V 至 +8V 時。對於具有 CIM 功能的記憶體裝置, 讀取偏壓約為 2.8V, 這可更進一步降低讀取干擾。

【0032】第 4D 圖顯示在本案一實施例中, 記憶體裝置具有小的隨機電報雜訊(Random Telegraph Noise (RTN))。當單元電流 I_{cell} 為 150nA 時, 隨機電報雜訊僅為 $\pm 0.02 \mu A$, 相當於平均值的 1.9% 左右。

【0033】此外, 在本案一實施例中, 可將操作電壓 VCC(例如為 3.3V)降壓而產生字元線電壓(例如但不受限於, 為 2.8V), 故而無需額外的電荷幫浦, 具有節省成本的優點。

【0034】在本案一實施例中，3D 記憶體裝置可提供平行的 $N \times M$ MAC 運算，以提供高運算頻寬。

【0035】此外，在本案一實施例中，如果可以提供更多個字元線電壓與更多個 ADC 輸出的話，則可大幅提高運算量。例如，如果可以提供 1000 個字元線電壓($N=1000$)與 1000 個 ADC 輸出($M=1000$)的話(約等於 8Mb 的記憶體磚(memory tile))，則可在極短讀取時間(如 150ns，8 位元 ADC 的輸出讀取約為 150ns)計算高達 1M 的 MAC 運算量，這相當於 6.7TOPS 的 MAC 運算能力，其中，TOPS 是 Tera Operations Per Second(每秒兆次運算)的縮寫，1 個 TOPS 代表每秒鐘可進行一兆次(10^{12})操作。

【0036】更一步說，本案實施例在佔用極小記憶體電路面積下，可進行超高速 MAC 運算。故而，本案實施例具有低電路成本但高速運算的優點。

【0037】綜上所述，雖然本發明已以實施例揭露如上，然其並非用以限定本發明。本發明所屬技術領域中具有通常知識者，在不脫離本發明之精神和範圍內，當可作各種之更動與潤飾。因此，本發明之保護範圍當視後附之申請專利範圍所界定者為準。

【符號說明】

【0038】

100: 記憶體裝置

110: 記憶體陣列

D_LBL(1)~ D_LBL(M): 區域位元線解碼器

D_{LSL}(1)~ D_{LSL}(M): 區域源極線解碼器

ADC(1)~ADC(M): 轉換單元

BLT(1)~BLT(Q): 位元線電晶體

SLT(1)~SLT(Q): 源極線電晶體

MC(i, j, k): 記憶體單元

WL(1)~WL(N): 字元線

LBL: 區域位元線

LSL: 區域源極線

310~350: 步驟

【發明申請專利範圍】

【請求項1】一種記憶體裝置之操作方法，該操作方法包括：
於進行一乘積累加運算(Multiply Accumulate, MAC)操作時，透過複數個第一信號線輸入複數個輸入至該記憶體裝置之複數個記憶體單元；

根據該些記憶體單元的複數個權重，該些記憶體單元輸出複數個單元電流於複數個第二信號線；

加總各該些第二信號線上的該些單元電流成複數個信號線電流；

加總該些信號線電流成一整體信號線電流；以及

由一轉換單元將該整體信號線電流轉換成一輸出，其中，該輸出代表該些輸入與該些權重的一乘積累加運算操作結果，其中，於進行 x 階權重運算時，耦接至該轉換單元的該些第二信號線的數量為 $Q=2^x-1$ ，其中， x 與 Q 皆為正整數，以及，該些記憶體單元為一單階儲存單元。

【請求項2】一種記憶體裝置，包括：

一記憶體陣列，包括複數個記憶體單元，該些記憶體單元儲存複數個權重，該些記憶體單元耦接至複數個第一信號線與複數個第二信號線；

至少一第一區域信號線解碼器，耦接至該記憶體陣列與至少一第一整體信號線；以及

至少一轉換單元，耦接至該至少一第一區域信號線解碼器與該至少一第一整體信號線，

其中，

於進行一乘積累加運算(Multiply Accumulate, MAC)操作時，複數個輸入透過該些第一信號線輸入至該些記憶體單元；根據該些記憶體單元的該些權重，該些記憶體單元輸出複數個單元電流於該些第二信號線；

該些單元電流在各該些第二信號線上加總成複數個信號線電流並輸入至該至少一第一區域信號線解碼器；

該至少一第一區域信號線解碼器將該些信號線電流加總成一整體信號線電流；以及

該至少一第一轉換單元將該至少一第一區域信號線解碼器所輸出的該整體信號線電流轉換以得到一輸出，其中，該輸出代表該些輸入與該些權重的一乘積累加運算操作結果，

其中，於進行 x 階權重運算時，耦接至該至少一轉換單元的該些第二信號線的數量為 $Q=2^x-1$ ，其中， x 與 Q 皆為正整數，以及，

該些記憶體單元為一單階儲存單元。

【請求項3】如請求項2所述之記憶體裝置，其中，該記憶體裝置為一三維AND記憶體裝置。

【請求項4】如請求項2所述之記憶體裝置，其中，該至少一第一整體信號線包括複數個第一整體信號線，該至少一轉換單元

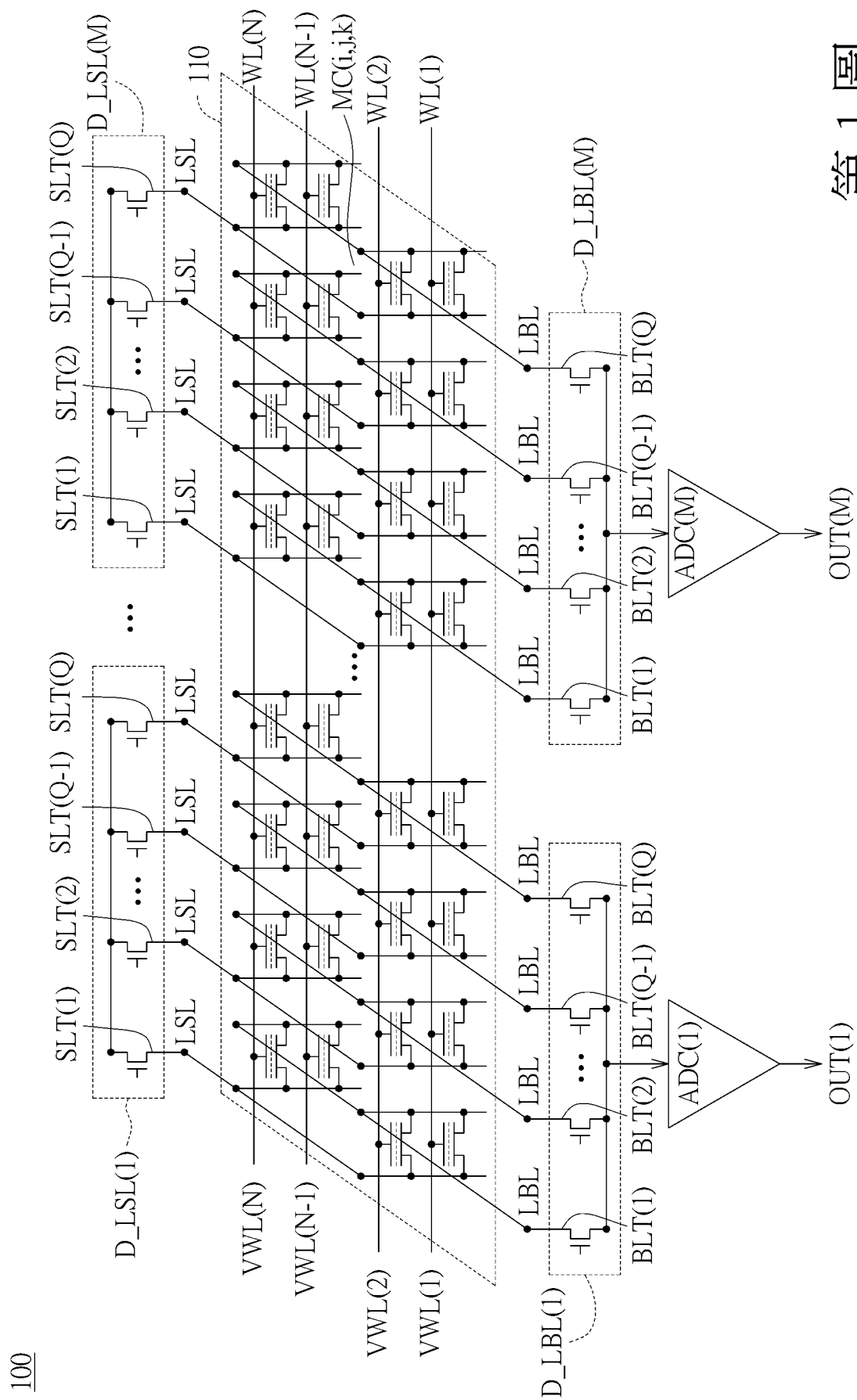
包括複數個轉換單元，各該些第一整體信號線耦接至各該些轉換單元。

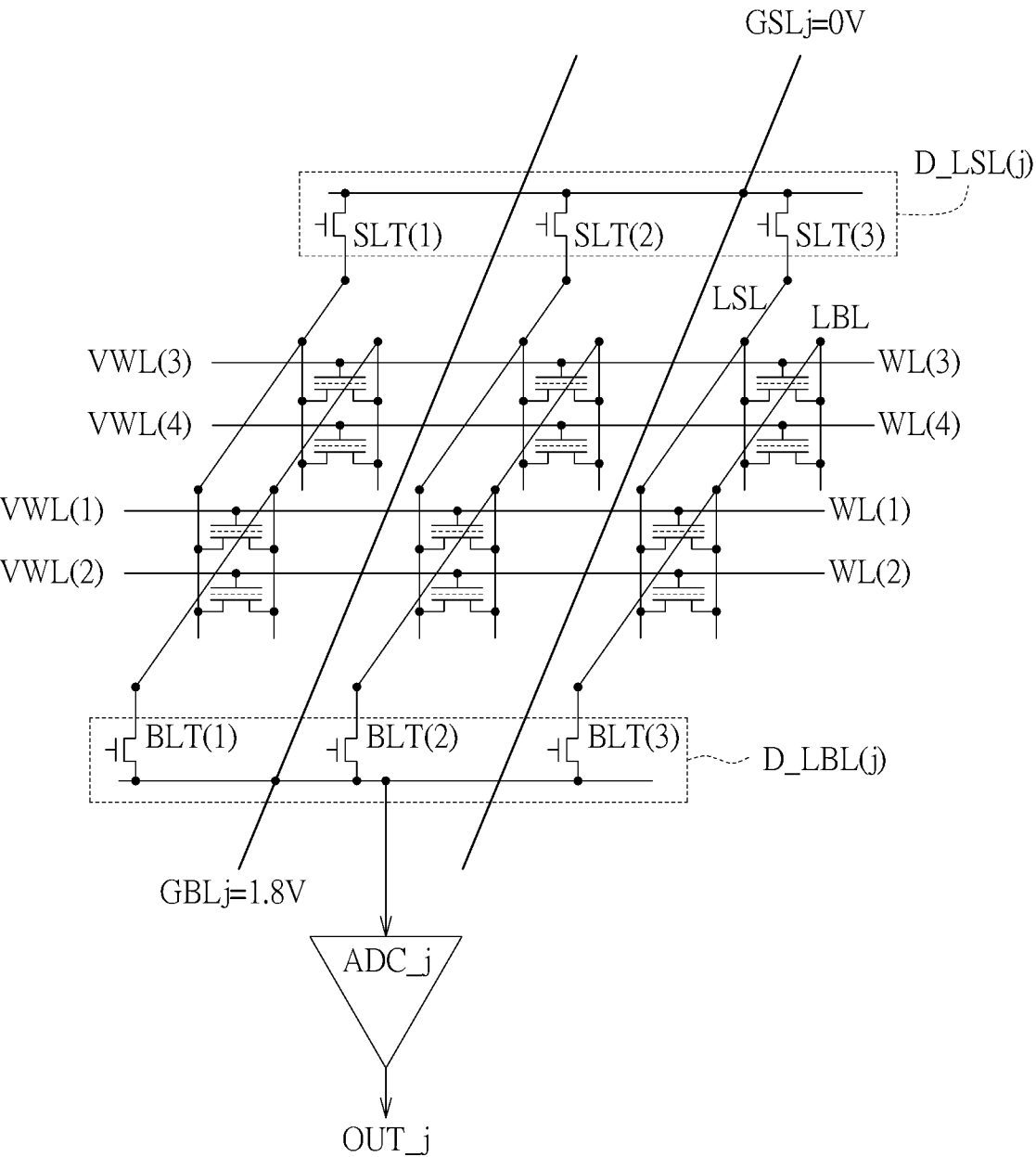
【請求項5】如請求項2所述之記憶體裝置，其中，於進行該乘積累加運算操作時，該至少一第一區域信號線解碼器的複數個第一電晶體為導通。

【請求項6】如請求項5所述之記憶體裝置，其中，該第一電晶體具有：一第一端耦接至該些第二信號線之一，一第二端耦接至該至少一轉換單元之一輸入端與該至少一第一整體信號線，以及一控制端接收一控制信號。

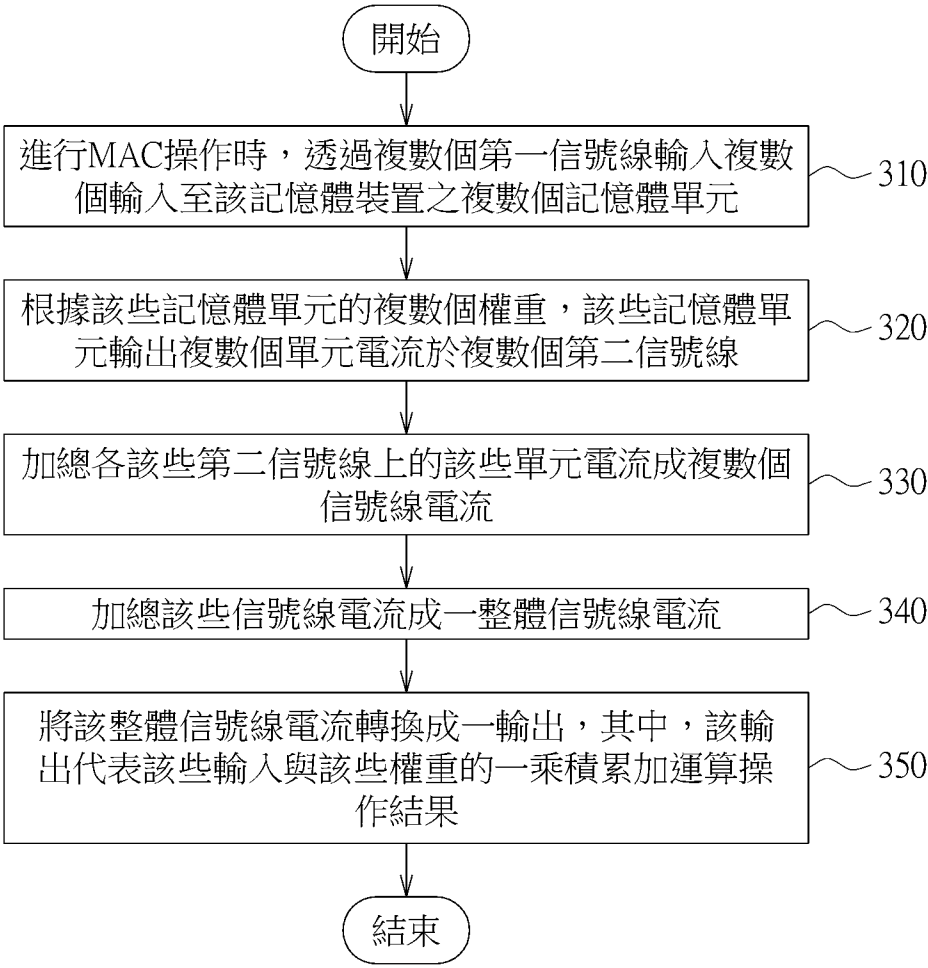
【請求項7】如請求項2所述之記憶體裝置，其中，該至少一轉換單元為一類比數位轉換單元(ADC)。

【請求項8】如請求項2所述之記憶體裝置，更包括至少一第二區域信號線解碼器，該至少一第二區域信號線解碼器包括複數個第二電晶體，該第二電晶體具有：一第一端耦接至複數個第三信號線之一，一第二端耦接至至少一第二整體信號線，以及一控制端接收一控制信號，其中，該些第三信號線耦接至該些記憶體單元。

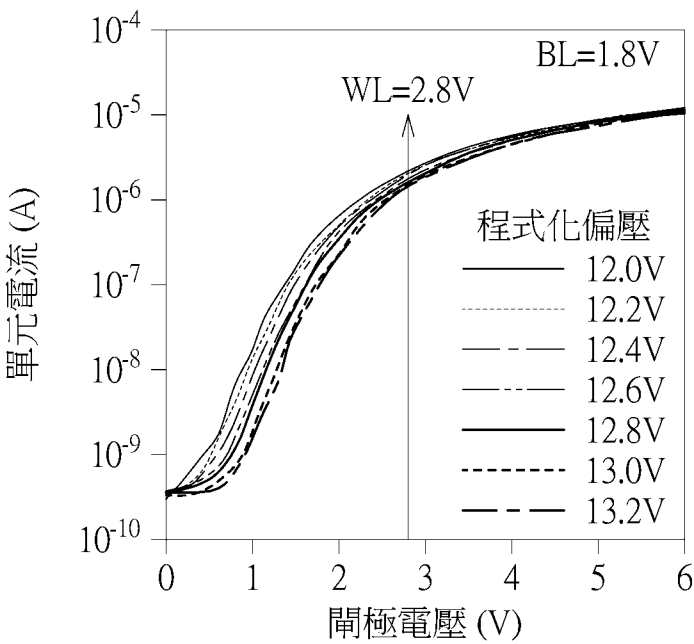




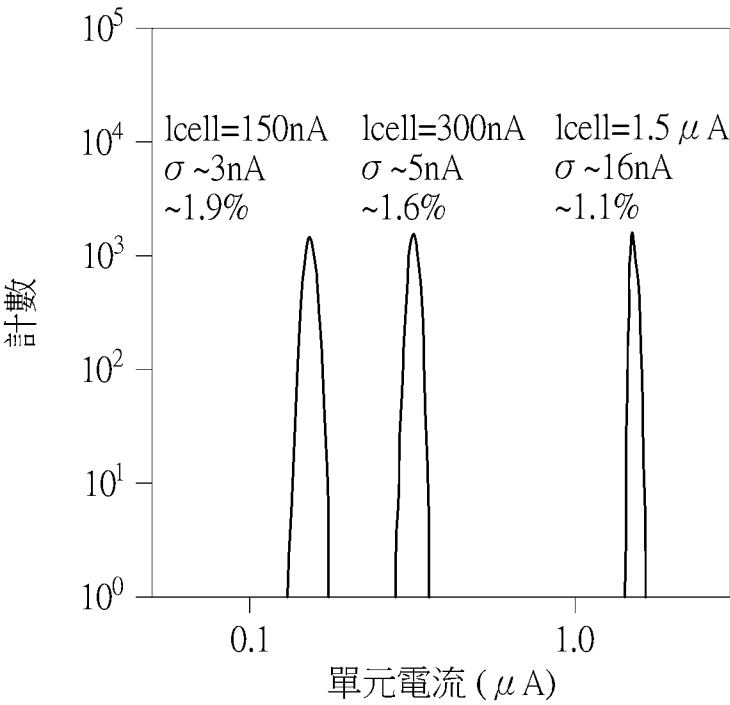
第 2 圖



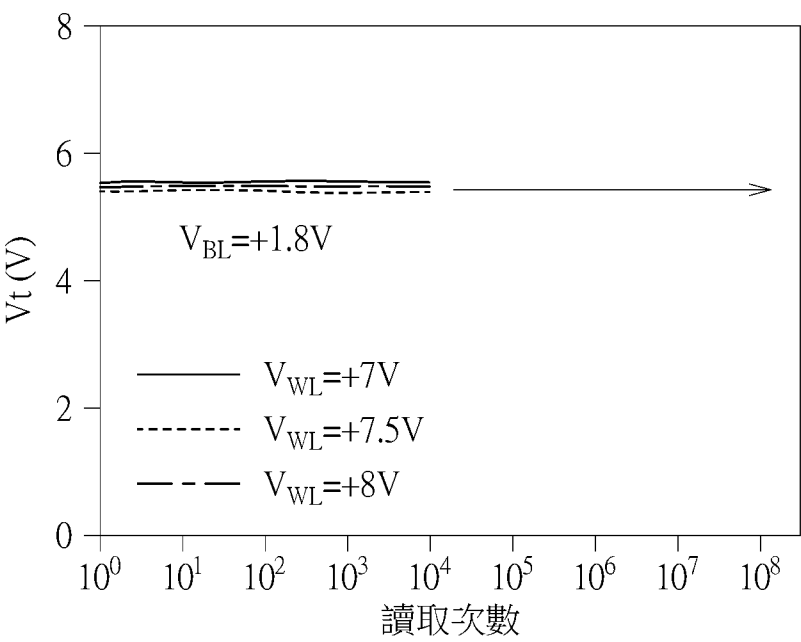
第 3 圖



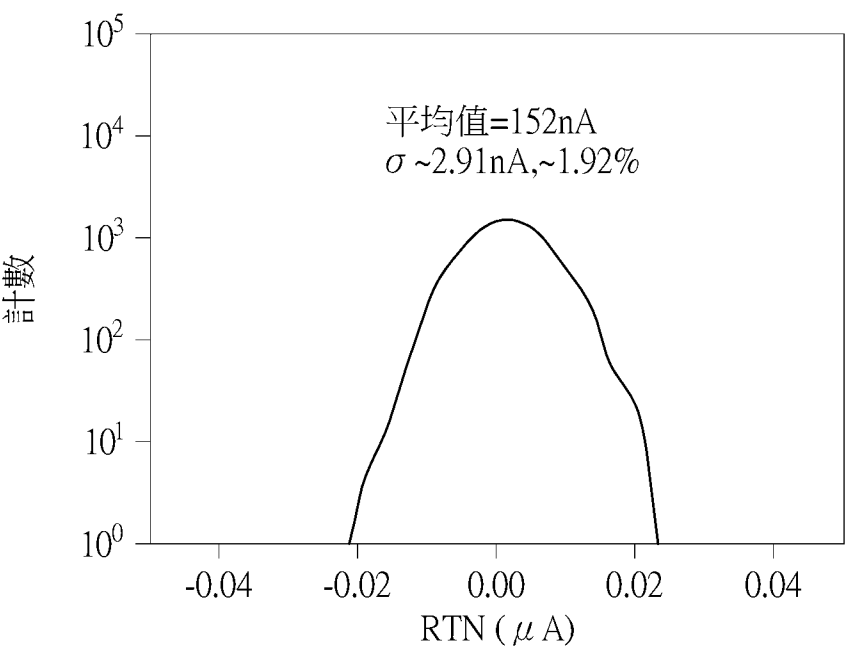
第 4A 圖



第 4B 圖



第 4C 圖



第 4D 圖