



(19)中華民國智慧財產局

(12)發明說明書公開本

(11)公開編號：TW 202321952 A

(43)公開日：中華民國 112 (2023) 年 06 月 01 日

(21)申請案號：111110903

(22)申請日：中華民國 111 (2022) 年 03 月 23 日

(51)Int. Cl. : G06F17/16 (2006.01)

G06N3/06 (2006.01)

G11C11/54 (2006.01)

(30)優先權：2021/11/22 美國

63/281,734

(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：胡瀚文 HU, HAN-WEN (TW)；李永駿 LI, YUNG-CHUN (TW)；林柏榕 LIN, BO-RONG (TW)；王淮慕 WANG, HUI-MU (TW)；王韋程 WANG, WEI-CHEN (TW)

(74)代理人：祁明輝；林素華

申請實體審查：有 申請專利範圍項數：10 項 圖式數：8 共 38 頁

(54)名稱

記憶體裝置及其操作方法

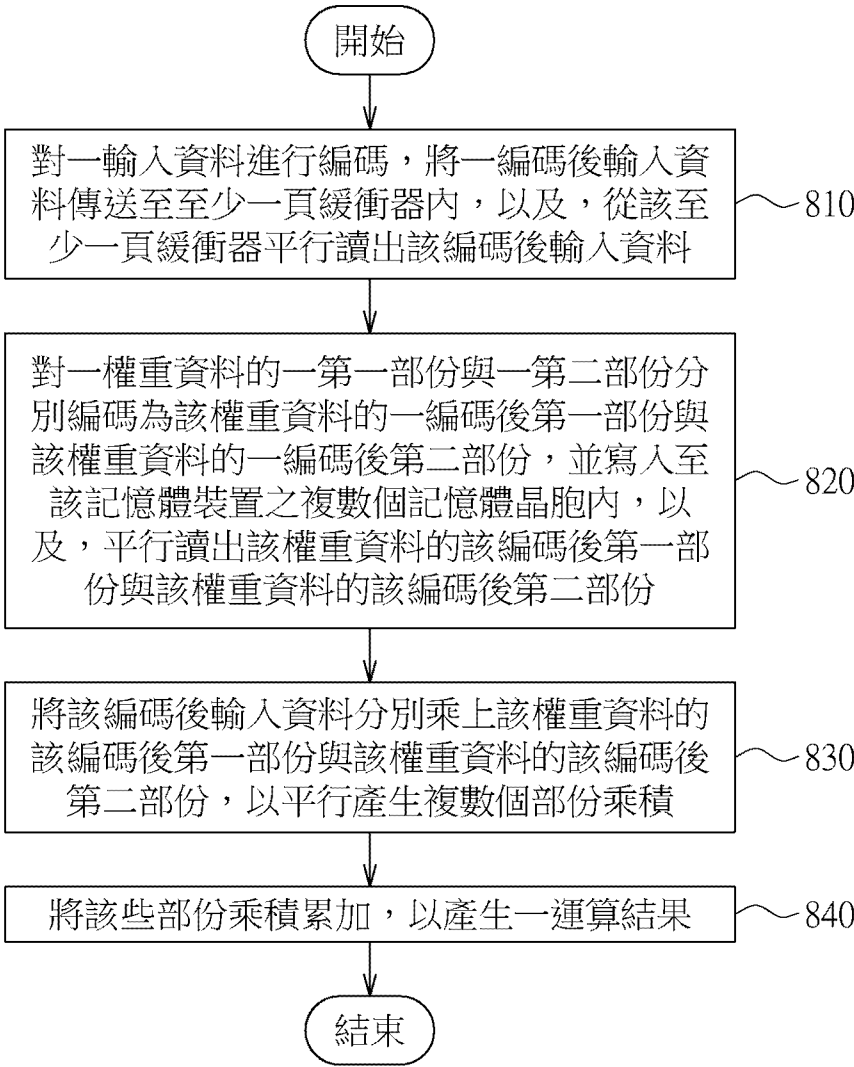
(57)摘要

本案提供記憶體裝置與其操作方法。記憶體裝置之操作方法包括：對一輸入資料進行編碼，將一編碼後輸入資料傳送至至少一頁緩衝器內，以及，從該至少一頁緩衝器平行讀出該編碼後輸入資料；對一權重資料的一第一部份與一第二部份分別編碼為該權重資料的一編碼後第一部份與該權重資料的一編碼後第二部份，並將該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份寫入至該記憶體裝置之複數個記憶體晶胞內，以及，平行讀出該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份；將該編碼後輸入資料分別乘上該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份，以平行產生複數個部份乘積；以及將該些部份乘積累加，以產生一運算結果。

A memory device and an operation method thereof are provided. A memory device and an operation method thereof are provided. The operation method includes: encoding an input data, sending an encoded input data to at least one page buffer, and reading out the encoded input data in parallel; encoding a first part and a second part of a weight data into an encoded first part and an encoded second part of the weight data, respectively, writing the encoded first part and the encoded second part of the weight data into a plurality of memory cells of the memory device, and reading out the encoded first part and the encoded second part of the weight data in parallel; multiplying the encoded input data with the encoded first part and the encoded second part of the weight data respectively to parallel generate a plurality of partial products; and accumulating the partial products to generate an operation result.

指定代表圖：

符號簡單說明：
810-840:步驟



第 8 圖

【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】

本案提供記憶體裝置與其操作方法。記憶體裝置之操作方法包括：對一輸入資料進行編碼，將一編碼後輸入資料傳送至至少一頁緩衝器內，以及，從該至少一頁緩衝器平行讀出該編碼後輸入資料；對一權重資料的一第一部份與一第二部份分別編碼為該權重資料的一編碼後第一部份與該權重資料的一編碼後第二部份，並將該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份寫入至該記憶體裝置之複數個記憶體晶胞內，以及，平行讀出該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份；將該編碼後輸入資料分別乘上該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份，以平行產生複數個部份乘積；以及將該些部份乘積累加，以產生一運算結果。

【英文】

A memory device and an operation method thereof are provided. A memory device and an operation method thereof are provided. The operation method includes: encoding an input data, sending an encoded input data to at least one page buffer, and reading out the encoded input data in parallel; encoding a first part and a second part of a weight data into an encoded first part

and an encoded second part of the weight data, respectively, writing the encoded first part and the encoded second part of the weight data into a plurality of memory cells of the memory device, and reading out the encoded first part and the encoded second part of the weight data in parallel; multiplying the encoded input data with the encoded first part and the encoded second part of the weight data respectively to parallel generate a plurality of partial products; and accumulating the partial products to generate an operation result.

【指定代表圖】 第 8 圖。

【代表圖之符號簡單說明】

810-840:步驟

【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】 本發明係有關於一種具有記憶體內運算(In-Memory-Computing (IMC))的記憶體裝置及其操作方法。

【先前技術】

【0002】 人工智慧(AI)已在許多領域中成為高度有效解決方案。AI的關鍵操作在於對大量的輸入資料(如輸入特徵圖(input feature maps))與權重值進行乘積累加運算(multiply-and-accumulate, MAC)。

【0003】 然而，以目前的AI架構而言，容易遇到輸出入瓶頸(IO bottleneck)與低效率的MAC運算流程。

【0004】 為達到高準確度，可執行具有多位元輸入及多位元權重值的MAC操作。然而，輸出入瓶頸變得更加嚴重，且效率將更低。

【0005】 記憶體內運算(In-Memory-Computing (IMC))可用於加速MAC運算，因為IMC可減少在中央處理架構下所需要用的複雜算術邏輯單元(Arithmetic logic unit, ALU)，且提供記憶體內的MAC操作的高並行性(parallelism)。

【0006】在進行IMC時，無符號數(unsigned integer)乘法運算與帶符號數(signed integer)乘法運算之說明如下。

【0007】例如，想要相乘兩個無符號數(皆為8位元): $a[7:0]$ 與 $b[7:0]$ 。則可執行8次的單位元乘法來產生8個部份乘積(partial product) $p0[7:0] \sim p7[7:0]$ ，各該8個部份乘積相關於被乘數a的各位元，該8個部份乘積可表示如下：

$$p0[7:0] = a[0] \times b[7:0] = \{8\{a[0]\}\} \& b[7:0]$$

$$p1[7:0] = a[1] \times b[7:0] = \{8\{a[1]\}\} \& b[7:0]$$

$$p2[7:0] = a[2] \times b[7:0] = \{8\{a[2]\}\} \& b[7:0]$$

$$p3[7:0] = a[3] \times b[7:0] = \{8\{a[3]\}\} \& b[7:0]$$

$$p4[7:0] = a[4] \times b[7:0] = \{8\{a[4]\}\} \& b[7:0]$$

$$p5[7:0] = a[5] \times b[7:0] = \{8\{a[5]\}\} \& b[7:0]$$

$$p6[7:0] = a[6] \times b[7:0] = \{8\{a[6]\}\} \& b[7:0]$$

$$p7[7:0] = a[7] \times b[7:0] = \{8\{a[7]\}\} \& b[7:0]$$

【0008】其中， $\{8\{a[0]\}\}$ 代表將 $a[0]$ 重複8次，其餘可依此類推。

【0009】為得到乘積，將該8個部份乘積 $p0[7:0] \sim p7[7:0]$ 相加，如第1A圖所示。第1A圖顯示兩個無符號數(皆為8位元)之相乘。

【0010】其中， $P0 = p0[0] + 0 + 0 + 0 + 0 + 0 + 0 + 0$ ，而 $P1 = p0[1] + p1[0] + 0 + 0 + 0 + 0 + 0 + 0$ ，其餘可依此類推。

【0011】乘積 $P[15:0]$ 則是將 $P0\sim P15$ 而得。乘積 $P[15:0]$ 代表將兩個無符號數(皆為8位元)相乘所得到的16位元無符號乘積。

【0012】而如果 b 是帶符號數，則於加總之前，部份乘積需要做符號展開(*sign-extended*)至乘積寬度。如果 a 也是帶符號數，則部份乘積 $P7$ 要從最後總和減去，而不是相加。

【0013】第1B圖顯示兩個符號數(皆為8位元)之相乘。於第1B圖中，符號「 $\sim$ 」代表互補，例如， $\sim p1[7]$ 代表 $p1[7]$ 的互補數。

【0014】在進行IMC時，如果能加快「操作速度」及減少容量需要(*capacity requirement*)的話，對於IMC性能將可有所助益。

【發明內容】

【0015】根據本案一實例，提出一種記憶體裝置，包括：複數個記憶體晶粒，各該些記憶體晶粒包括複數個記憶體平面、複數個頁緩衝器與一累加電路，各該些記憶體平面包括複數個記憶體晶胞。其中，對一輸入資料進行編碼，將一編碼後輸入資料傳送至至少一頁緩衝器內，以及，從該至少一頁緩衝器平行讀出該編碼後輸入資料；對一權重資料的一第一部份與一第二部份分別編碼為該權重資料的一編碼後第一部份與該權重資料的一編碼後第二部份，並寫入至該記憶體裝置之該些記憶體晶胞內，以及，平行讀出該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份；將該編碼後輸入資料分別乘上該權重資料的該編碼

後第一部份與該權重資料的該編碼後第二部份，以平行產生複數個部份乘積；以及將該些部份乘積累加，以產生一運算結果。

【0016】根據本案另一實例，提出一種記憶體裝置之操作方法，包括：對一輸入資料進行編碼，將一編碼後輸入資料傳送至至少一頁緩衝器內，以及，從該至少一頁緩衝器平行讀出該編碼後輸入資料；對一權重資料的一第一部份與一第二部份分別編碼為該權重資料的一編碼後第一部份與該權重資料的一編碼後第二部份，並將該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份寫入至該記憶體裝置之複數個記憶體晶胞內，以及，平行讀出該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份；將該編碼後輸入資料分別乘上該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份，以平行產生複數個部份乘積；以及將該些部份乘積累加，以產生一運算結果。

【0017】為了對本發明之上述及其他方面有更佳的了解，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】

【0018】

第1A圖顯示兩個無符號數之相乘。

第1B圖顯示兩個符號數之相乘。

第2圖顯示根據本案一實施例之記憶體裝置之操作方法流程圖。

第3A圖與第3B圖顯示本案實施例中之錯誤位元容忍資料編

碼。

第4A圖顯示在本案一實施例中的8位元無符號數乘法運算，而第4B圖顯示在本案一實施例中的8位元帶符號數乘法運算。

第5A圖顯示根據本案一實施例之無符號數乘法運算之操作示意圖。第5B圖顯示根據本案一實施例之帶符號數乘法運算之操作示意圖。

第6圖顯示根據本案一實施例之記憶體裝置之功能方塊圖。

第7圖顯示比較本案一實施例與現有技術之MAC運算流程時序圖。

第8圖顯示根據本案一實施例之記憶體裝置之操作方法。

【實施方式】

【0019】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。

【0020】第2圖顯示根據本案一實施例之記憶體裝置之操作方法流程圖。於步驟210中，對輸入資料進行編碼並將編碼後輸入資料(其為向量)以傳送至複數個頁緩衝器內，且從該些頁緩衝器平行讀出。至於如何對輸入資料進行編碼的細節將於底下說明

之。

【0021】於步驟 220 中，對權重資料進行編碼並將編碼後權重資料(其為向量)寫入至記憶體裝置之複數個記憶體晶胞內，以及，平行讀出編碼後權重資料。至於如何對權重資料進行編碼的細節將於底下說明之。其中，在進行編碼時，權重資料的最高有效位元(most significant bit，MSB)與最低有效位元(least significant bit，LSB)被分別編碼。

【0022】於步驟 230 中，將輸入資料分別乘上編碼後權重資料的 MSB 與 LSB，以平行產生複數個部份乘積(partial product)。

【0023】於步驟 240 中，將該些部份乘積加總(累積)，以產生乘積累加運算(multiply-and-accumulate，MAC)結果或者漢明距離(Hamming distance)運算結果。

【0024】本案一實施例揭露可實施數位 MAC 操作的記憶體裝置，具有錯誤位元容忍(error-bit-tolerance)資料編碼，以容忍錯誤位元及減少面積需求。錯誤位元容忍資料編碼乃是使用輸入資料複製(duplication)及權重資料平坦化(flattening)技術。此外，本案實施例的感應技術包括標準的單位元晶胞(single level cell，SLC)讀取與邏輯及(AND)功能，進行位元乘法以產生部份乘積。於本案其他可能實施例中，如果在感應過程中，頁緩衝器不會將存於閃鎖單元內的輸入資料移除的話，則 SLC 讀取可由選擇位元讀取所取代，或者被多位元晶胞(Multi-Level Cell，MLC)、三

位元晶胞(Triple Level Cell, TLC)、四位元晶胞(Quad-level cells, QLC)讀取操作所取代。此外，本案一實施例的數位 MAC 操作運算乃是使用高頻寬權重累加器(high bandwidth weighted accumulator)以產生輸出結果，該高頻寬權重累加器可藉由重複使用故障位元計數(fail-bit-count, FBC)電路來實施權重化累加(weighted accumulation)。

【0025】 本案另一實施例揭露一種可實施漢明距離運算的記憶體裝置，具有錯誤位元容忍資料編碼，以容忍錯誤位元。錯誤位元容忍資料編碼乃是使用輸入資料複製及權重資料平坦化技術。此外，本案實施例的感應技術包括標準的單位元晶胞(single level cell, SLC)讀取與邏輯互斥或(EXOR)功能，進行位元乘法以產生部份乘積。於本案其他可能實施例中，如果在感應過程中，頁緩衝器不會將存於閃鎖單元內的輸入資料移除的話，則 SLC 讀取可由選擇位元讀取所取代，或者被 MLC、TLC、QLC 讀取操作所取代。而邏輯互斥或(EXOR)功能可由邏輯反互斥或(XNOR)與邏輯反所取代。此外，本案一實施例的數位漢明距離操作運算乃是使用高頻寬未權重累加器(high bandwidth unweighted accumulator)以產生輸出結果，該高頻寬未權重累加器可藉由重複使用故障位元計數(fail-bit-count, FBC)電路來實施未權重累加(unweighted accumulation)。

【0026】 第 3A 圖與第 3B 圖顯示本案實施例中之錯誤位元容忍資料編碼。例如但不受限於，輸入資料與權重資料為浮點

32(floating point 32)資料。於第 3A 圖中，將輸入資料與權重資料量化成 8 位元二進位整數，其中，輸入資料與權重資料皆為 8 位元向量，且為 N 維度(N 為正整數)。輸入資料與權重資料可分別表示為 $X_i(7:0)$ 與 $W_i(7:0)$ 。

【0027】於第 3B 圖中，將該些 N 維度的各個 8 位元權重向量分開成 MSB 向量與 LSB 向量。8 位元權重向量的 MSB 向量包括 4 位元 $W_i(7:4)$ ，而 LSB 向量包括 4 位元 $W_i(3:0)$ 。

【0028】接著，將 8 位元權重向量的 MSB 向量與 LSB 向量的各位元以一元編碼(Unary coding)(亦即數值形式(value format))表示。例如，8 位元權重向量的 MSB 向量的位元 $W_{i=0}(7)$ 可以表示為 8 位元(複製 8 次)，8 位元權重向量的 MSB 向量的位元 $W_{i=0}(6)$ 可以表示為 4 位元(複製 4 次)，8 位元權重向量的 MSB 向量的位元 $W_{i=0}(5)$ 可以表示為 2 位元(複製 2 次)，8 位元權重向量的 MSB 向量的位元 $W_{i=0}(4)$ 可以表示為 1 位元(複製 1 次)，並且將備用位元(spare bit)(0)加入於 8 位元權重向量的 MSB 向量的位元 $W_{i=0}(4)$ 之後。如此，可將 8 位元權重向量的 4 位元 MSB 向量編碼成 16 位元的一元編碼(Unary coding)格式。

【0029】同樣地，可將 8 位元權重向量的 4 位元 LSB 向量編碼成 16 位元的一元編碼(Unary coding)格式。

【0030】在本案一實施例中，透過上次的編碼方式，可以提高錯誤位元的容忍度。

【0031】第 4A 圖顯示在本案一實施例中的 8 位元無符號數

(unsigned integer)乘法運算，而第 4B 圖顯示在本案一實施例中的 8 位元帶符號數(signed integer)乘法運算。

【0032】如第 4A 圖所示，於進行 8 位元無符號數乘法運算時，在第 0 周期時，將輸入資料的 $X_i(7)$ (輸入資料已編碼成一元編碼格式)乘上權重資料的 MSB 向量 $W_i(7:4)$ (權重資料的 MSB 向量已編碼成一元編碼格式)，以得到第一 MSB 部份乘積。相似地，將輸入資料的 $X_i(7)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ (權重資料的 LSB 向量已編碼成一元編碼格式)，以得到第一 LSB 部份乘積。將第一 MSB 部份乘積位移 4 位元後相加至第一 LSB 部份乘積，以得到第一部份乘積。

【0033】在第 1 周期時，將輸入資料的 $X_i(6)$ 乘上權重資料的 MSB 向量 $W_i(7:4)$ (權重資料的 MSB 向量已編碼成一元編碼格式)，以得到第二 MSB 部份乘積。相似地，將輸入資料的 $X_i(6)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ (權重資料的 LSB 向量已編碼成一元編碼格式)，以得到第二 LSB 部份乘積。將第二 MSB 部份乘積位移 4 位元後相加至第二 LSB 部份乘積以得到第二部份乘積。此外，更將第一部份乘積位移 1 位元相加至第二部份乘積，以得到更新後第二部份乘積。其餘周期(第 2 周期至第 7 周期)的操作可依此類推，於此不重述。

【0034】亦即，透過 8 個周期可以完成 8 位元無符號數乘法運算。

【0035】如第 4B 圖所示，於進行 8 位元帶符號數乘法運算時，

在第 0 周期時，將輸入資料的 $X_i(7)$ 乘上權重資料的 MSB 向量 $W_i(7)$ (權重資料的 MSB 向量已編碼成一元編碼格式)，以及將輸入資料的 $X_i(7)$ 乘上權重資料的 MSB 向量 $W_i(6:4)$ (權重資料的 MSB 向量已編碼成一元編碼格式) 並反相，兩者相加，以得到第一 MSB 部份乘積。將輸入資料的 $X_i(7)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ (權重資料的 LSB 向量已編碼成一元編碼格式) 並反相以得到第一 LSB 部份乘積。將第一 MSB 部份乘積位移 4 位元後相加至第一 LSB 部份乘積，以得到第一部份乘積。

【0036】 在第 1 周期時，將輸入資料的 $X_i(6)$ 乘上權重資料的 MSB 向量 $W_i(7)$ (權重資料的 MSB 向量已編碼成一元編碼格式) 後反相，以及，將輸入資料的 $X_i(6)$ 乘上權重資料的 MSB 向量 $W_i(6:4)$ (權重資料的 MSB 向量已編碼成一元編碼格式)，兩者相加，以得到第二 MSB 部份乘積。相似地，將輸入資料的 $X_i(6)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ (權重資料的 LSB 向量已編碼成一元編碼格式)，以得到第二 LSB 部份乘積。將第二 MSB 部份乘積位移 4 位元後相加至第二 LSB 部份乘積以得到第二部份乘積。此外，更將第一部份乘積位移 1 位元相加至第二部份乘積，以得到更新後第二部份乘積。其餘周期(第 2 周期至第 7 周期)的操作可依此類推，於此不重述。

【0037】 亦即，透過 8 個周期可以完成 8 位元帶符號數乘法運算。

【0038】 上述方式需要 8 個周期才可以完成 8 位元無符號數

乘法運算與 8 位元帶符號數乘法運算。

【0039】第 5A 圖顯示根據本案一實施例之無符號數乘法運算之操作示意圖。第 5B 圖顯示根據本案一實施例之帶符號數乘法運算之操作示意圖。第 5A 圖與第 5B 圖以輸入資料與權重資料都是 8 位元為例做說明，但當知本案並不受限於此。

【0040】於第 5A 圖與第 5B 圖中，輸入資料亦被編碼，以及，權重資料的 MSB 向量與 LSB 向量已編碼成一元編碼格式。

【0041】於第 5A 圖與第 5B 圖中，將輸入資料輸入至頁緩衝器，而權重資料則是寫入至複數個記憶體晶胞內。

【0042】在第 5A 圖中，輸入資料從頁緩衝器平行讀取出，以及，將權重資料從該些記憶體晶胞平行讀取出，進行平行乘法，以得到部份乘積。

【0043】細言之，輸入資料的位元 $X_i(7)$ 乘上權重資料的 MSB 向量 $W_i(7:4)$ ，以得到第一 MSB 部份乘積。輸入資料的位元 $X_i(6)$ 乘上權重資料的 MSB 向量 $W_i(7:4)$ ，以得到第二 MSB 部份乘積。其餘可依此類推，直到輸入資料的位元 $X_i(0)$ 乘上權重資料的 MSB 向量 $W_i(7:4)$ ，以得到第八 MSB 部份乘積。例如，在第 5A 圖中，輸入資料的位元 $X_i(7)$ 被複製 15 次，並加上備用位元，以成為 16 位元的乘數“0000000000000000”。此 16 位元的乘數“0000000000000000”乘上權重資料的 MSB 向量 $W_i(7:4)$ “1111111100001100”，以得到第一 MSB 部份乘積“0000000000000000”。其餘可依此類推。所有的 MSB 部份乘積

可以合併成為輸入串流(input stream) M 。

【0044】相似地，將輸入資料的 $X_i(7)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ ，以得到第一 LSB 部份乘積。將輸入資料的 $X_i(6)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ ，以得到第二 LSB 部份乘積。其餘可依此類推，直到輸入資料的位元 $X_i(0)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ ，以得到第八 LSB 部份乘積。所有的 LSB 部份乘積可以合併成為輸入串流 L 。

【0045】之後，將該些第一至第八 MSB 部份乘積與該些第一至第八 LSB 部份乘積合併，並且計數合併值的位元 1 的數量，即可得到無符號數乘法運算之 MAC 運算結果。

【0046】在第 5B 圖中，輸入資料從頁緩衝器平行讀取出，以及，將權重資料從該些記憶體晶胞平行讀取出，進行平行乘法，以得到部份乘積。

【0047】細言之，輸入資料的位元 $X_i(7)$ 乘上權重資料的 MSB 向量 $W_i(7:4)$ ，以得到第一 MSB 部份乘積。輸入資料的位元 $X_i(6)$ 乘上權重資料的 MSB 向量 $W_i(7:4)$ ，以得到第二 MSB 部份乘積。其餘可依此類推，直到輸入資料的位元 $X_i(0)$ 乘上權重資料的 MSB 向量 $W_i(7:4)$ ，以得到第八 MSB 部份乘積。

【0048】相似地，將輸入資料的 $X_i(7)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ ，以得到第一 LSB 部份乘積。將輸入資料的 $X_i(6)$ 乘上權重資料的 LSB 向量 $W_i(3:0)$ ，以得到第二 LSB 部份乘積。其餘可依此類推，直到輸入資料的位元 $X_i(0)$ 乘上權重資料的

LSB 向量 $W_i(3:0)$ ，以得到第八 LSB 部份乘積。

【0049】之後，將該些第一至第八 MSB 部份乘積與該些第一至第八 LSB 部份乘積合併，並且計數合併值的位元 1 的數量，即可得到帶符號數乘法運算之 MAC 運算結果。

【0050】第 6 圖顯示根據本案一實施例之記憶體裝置之功能方塊圖。記憶體裝置 600 包括複數個記憶體晶粒(die)615。在第 6 圖中以記憶體裝置 600 包括 4 個記憶體晶粒 615 為例做說明，但當知本案並不受限於此。

【0051】記憶體晶粒 615 包括複數個記憶體平面(memory plane, MP)620、複數個頁緩衝器 625 與累加電路 630。在第 6 圖中以記憶體晶粒 615 包括 4 個記憶體平面 620 與 4 個頁緩衝器 625 為例做說明，但當知本案並不受限於此。記憶體平面 620 包括複數個記憶體晶胞(未示出)。權重資料存於該些記憶體晶胞內。

【0052】在各記憶體晶粒 615 內，累加電路 630 由該些記憶體平面 620 所共享，故而，累加電路 630 依序執行該些記憶體平面 620 的累加運算。此外，各記憶體晶粒 615 可以獨立執行本案實施例的上述數位 MAC 運算與數位漢明距離運算。

【0053】輸入資料可以透過複數條字元線而輸入至該些頁緩衝器 625 內。

【0054】頁緩衝器 625 包括感應電路 631、複數個門鎖單元 633-641、複數個邏輯閘 643 與 645。

【0055】感應電路 631 耦接至位元線 BL，以感應位元線 BL 上的電流。

【0056】閘鎖單元 633-641 例如但不受限於，分別為資料閘鎖器(data latch，DL)633、閘鎖器(L1)635、閘鎖器(L2)637、閘鎖器(L3)639 與共同資料閘鎖器(common data latch，CDL)641。閘鎖單元 633-641 例如但不受限於，為單位元閘鎖器。

【0057】資料閘鎖器 633 用以閘鎖權重資料，並將權重資料輸出至邏輯閘 643 與 645。

【0058】閘鎖器(L1)635 與閘鎖器(L3)639 乃是用於解碼。

【0059】閘鎖器(L2)637 用以閘鎖輸入資料，並將輸入資料輸出至邏輯閘 643 與 645。

【0060】共同資料閘鎖器 641 用以閘鎖由邏輯閘 643 或 645 所傳來的資料。

【0061】邏輯閘 643 與 645 例如但不受限於，分別為邏輯及閘，以及邏輯 XOR 閘。邏輯閘 643 對輸入資料與權重資料進行邏輯及運算，並將邏輯運算結果寫入至共同資料閘鎖器 641。邏輯閘 645 對輸入資料與權重資料進行邏輯 XOR 運算，並將邏輯運算結果寫入至共同資料閘鎖器 641。邏輯閘 643 與 645 分別受控於致能信號 AND_EN 與 XOR_EN。例如，當進行數位 MAC 運算時，邏輯閘 643 被致能信號 AND_EN 致能；以及，當進行數位漢明距離運算時，邏輯閘 645 被致能信號 XOR_EN 致能。

【0062】以第 5A 圖或第 5B 圖來做說明，輸入資料的位元 $X_i(7)$ 的 1 個位元係輸入至門鎖器(L2)637，而已編碼成一元編碼格式的權重資料的 MSB 向量 $W_i(7:4)$ 的 1 個位元係輸入至資料門鎖器 633。門鎖器(L2)637 的輸入資料與資料門鎖器 633 的權重資料則由邏輯閘 643 或 645 進行邏輯運算後，共同資料門鎖器 641 用以門鎖由邏輯閘 643 或 645 所傳來的資料。共同資料門鎖器 641 亦可視為是該位元線的資料輸出路徑。

【0063】累加電路 630 包括：部份乘積累加單元 651、單一維度乘積產生單元 653、第一多維度累加單元 655、第二多維度累加單元 657 與權重累加控制單元 659。

【0064】部份乘積累加單元 651 耦接至頁緩衝器 625，以接收由頁緩衝器 625 的複數個共同資料門鎖器 641 所傳來的複數個邏輯運算結果，來產生複數個部份乘積。

【0065】例如，以第 5A 圖或第 5B 圖來說，部份乘積累加單元 651 產生該些第一至第八 MSB 部份乘積以及該些第一至第八 LSB 部份乘積。

【0066】單一維度乘積產生單元 653 耦接至部份乘積累加單元 651，將部份乘積累加單元 651 所產生的該些部份乘積累加，以產生單一維度乘積。

【0067】例如，以第 5A 圖或第 5B 圖來說，單一維度乘積產生單元 653 則將部份乘積累加單元 651 所產生的該些第一至第八 MSB 部份乘積以及該些第一至第八 LSB 部份乘積累加，以產生

單一維度乘積。

【0068】例如，於第 0 周期產生第<0>維度乘積後，可於第 1 周期產生第<1>維度乘積，其餘依此類推。

【0069】第一多維度累加單元 655 耦接至單一維度乘積產生單元 653，將單一維度乘積產生單元 653 所產生的複數個單一維度乘積累加，以得到多維度乘積累加結果。

【0070】例如但不受限於，第一多維度累加單元 655 將單一維度乘積產生單元 653 所產生的第<0>至第<7>維度乘積累加，以得到 8 維度<0:7>乘積累加結果。接著，第一多維度累加單元 655 將單一維度乘積產生單元 653 所產生的第<8>至第<15>維度乘積累加，以得到另一個 8 維度<8:15>乘積累加結果。

【0071】第二多維度累加單元 657 耦接至第一多維度累加單元 655，將第一多維度累加單元 655 所產生的複數個多維度乘積累加結果進行累加，以得到輸出累加值。例如但不受限於，第二多維度累加單元 657 將第一多維度累加單元 655 所產生的 64 個 8 維度乘積累加結果進行累加，以得到 512 維度的輸出累加值。

【0072】權重累加控制單元 659 耦接至部份乘積累加單元 651、單一維度乘積產生單元 653、第一多維度累加單元 655。根據進行數位 MAC 運算操作或數位漢明距離運算操作，權重累加控制單元 659 被致能或失能。例如但不受限於，當進行數位 MAC 運算操作時，權重累加控制單元 659 被致能；以及，當進行數位漢明距離運算操作時，權重累加控制單元 659 被失能。當權重累

加控制單元 659 被致能時，權重累加控制單元 659 根據權重累加致能信號 WACC_EN 而輸出控制信號至部份乘積累加單元 651、單一維度乘積產生單元 653、第一多維度累加單元 655。

【0073】第 6 圖中的單一個頁緩衝器 620 乃是耦接至複數條位元線 BL。例如但不受限於，各頁緩衝器 620 耦接至 131072 條位元線 BL，每一個周期內選擇 128 條位元線 BL 上的資料結果給累加電路 630 進行累加。如此的話，需要 1024 個周期把 131072 條位元線 BL 上的資料送完。

【0074】此外於上述說明中，部份乘積累加單元 651 一次接收 128 位元，第一多維度累加單元 655 產生 8 維度乘積累加結果，而第二多維度累加單元 657 產生 512 維度的輸出累加值。但本案不受限於此。於另一可能實施例中，部份乘積累加單元 651 一次接收 64 位元(2 位元為 1 組)，第一多維度累加單元 655 產生 16 維度乘積累加結果，而第二多維度累加單元 657 產生 512 維度的輸出累加值。

【0075】第 7 圖顯示比較本案一實施例與現有技術之 MAC 運算流程時序圖。以第 7 圖來看，於輸入廣播(input broadcasting)時間內，接收輸入資料。之後，對於該輸入資料與該權重資料進行如上述方式的位元乘法與位元累加，以產生 MAC 運算操作結果。

【0076】於習知技術中，需要較長的操作時間。相反的，在本案實施例中，藉由平行乘法來產生(1)輸入向量與權重資料的

MSB 向量的部份乘積，以及(2)輸入向量與權重資料的 LSB 向量的部份乘積。如此可以在 1 個周期內完成無符號數乘法運算及/或帶符號數乘法運算。所以，本案實施例的操作速度快於習知技術的操作速度。

【0077】第 8 圖顯示根據本案一實施例之記憶體裝置之操作方法，包括：對一輸入資料進行編碼，將一編碼後輸入資料傳送至至少一頁緩衝器內，以及，從該至少一頁緩衝器平行讀出該編碼後輸入資料(810)；對一權重資料的一第一部份與一第二部份分別編碼為該權重資料的一編碼後第一部份與該權重資料的一編碼後第二部份，並寫入至該記憶體裝置之複數個記憶體晶胞內，以及，平行讀出該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份(820)；將該編碼後輸入資料分別乘上該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份，以平行產生複數個部份乘積(830)；以及將該些部份乘積累加，以產生一運算結果(840)。

【0078】如上述般，於本案實施例中，透過位元錯誤容忍編碼方式可以減少錯誤位元，提高準確度並減少對記憶體容量的需求。

【0079】此外，本案一實施例的數位 MAC 操作運算乃是使用高頻寬權重累加器以產生輸出結果，該高頻寬權重累加器可藉由重複使用故障位元計數電路來實施權重化累加，故而可以改善累加速度。

【0080】 本案一實施例的數位漢明距離操作運算乃是使用高頻寬未權重累加器以產生輸出結果，該高頻寬未權重累加器可藉由重複使用故障位元計數電路來實施未權重累加，故而可以改善累加速度。

【0081】 本案上述實施例可應用於 NAND 型快閃記憶體，或者敏感於錯誤位元的記憶體裝置，例如但不受限於，NOR 型快閃記憶體，相變(PCM)型快閃記憶體，磁式隨機存取記憶體(magnetic RAM)或電阻式 RAM。

【0082】 於上述實施例中，累加電路 630 可以接收由頁緩衝器 625 所傳來的 128 個部份乘積，但於本案其他實施例中，累加電路 630 可以接收由頁緩衝器 625 所傳來的 2、4、8、16...512 個部份乘積(為 2 的冪次方)，此亦在本案精神範圍內。

【0083】 於上述實施例中，累加電路 630 可以支援加法功能，但於本案其他實施例中，累加電路 630 可以支援減法功能，此亦在本案精神範圍內。

【0084】 於上述實施例中，雖以 INT8 或 UNIT8 的 MAC 運算為例做說明，但於本案其他實施例中，也可支援 INT2、UNIT2、INT4、UNIT4 的 MAC 運算操作，此亦在本案精神範圍內。

【0085】 雖然上述實施例中，將權重資料分為 MSB 向量與 LSB 向量(2 個向量)，但本案並不受限於此。於本案其他可能實施例中，權重資料亦可分為更多個向量，此亦在本案精神範圍內。

【0086】 本案上述實施例可應用於需要 MAC 運算操作的 AI

模型設計中，例如但不受限於，完全連接層(fully-connection layer)、卷積層(convolution layer)、多層感知器(multiple layer Perceptron)、支援向量機器(support vector machine)等 AI 技術之中。

【0087】 本案上述不只可應用於計算用途(computing usage)，也可應用於相似性搜尋(similarity search)、分析用途(analysis usage)、聚類分析(clustering analysis)等。

【0088】 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】

【0089】

- 210-240:步驟
- 600:記憶體裝置
- 615:記憶體晶粒
- 620:記憶體平面
- 625: 頁緩衝器
- 630: 累加電路
- 631: 感應電路
- 633-641:閘鎖單元
- 643、645:邏輯閘
- 651:部份乘積累加單元
- 653:單一維度乘積產生單元
- 655:第一多維度累加單元
- 657:第二多維度累加單元
- 659:權重累加控制單元

810-840:步驟

【發明申請專利範圍】

【請求項1】一種記憶體裝置，包括：

複數個記憶體晶粒，各該些記憶體晶粒包括複數個記憶體平面、複數個頁緩衝器與一累加電路，各該些記憶體平面包括複數個記憶體晶胞，

其中，對一輸入資料進行編碼，將一編碼後輸入資料傳送至至少一頁緩衝器內，以及，從該至少一頁緩衝器平行讀出該編碼後輸入資料；

對一權重資料的一第一部份與一第二部份分別編碼為該權重資料的一編碼後第一部份與該權重資料的一編碼後第二部份，並寫入至該記憶體裝置之該些記憶體晶胞內，以及，平行讀出該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份；

將該編碼後輸入資料分別乘上該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份，以平行產生複數個部份乘積；以及

將該些部份乘積累加，以產生一運算結果。

【請求項2】如請求項1所述之記憶體裝置，其中，該權重資料的該第一部份是最高有效位元(most significant bit，MSB)，以及，該權重資料的該第二部份是最低有效位元(least significant bit，LSB)。

【請求項3】如請求項1所述之記憶體裝置，其中，

第1頁，共4頁(發明申請專利範圍)

於編碼時，將該輸入資料與該權重資料分別量化成二進位整數向量；

將該輸入資料的各位元複製複數次並加上一備用位元；

將該權重資料分開成該第一部份與該第二部份；以及

將該權重資料的該第一部份與該第二部份的各位元以一元編碼表示，以得到該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份。

【請求項4】如請求項1所述之記憶體裝置，其中，該運算結果包括一乘積累加結果 (multiply-and-accumulate, MAC) 運算結果或者一漢明距離 (Hamming distance) 運算結果；以及，將屬同一維度的該些部份乘積累加，以得到單一維度乘積；將複數個單一維度乘積累加，以得到一多維度乘積累加結果；將複數個多維度乘積累加結果累加，以產生該運算結果。

【請求項5】如請求項4所述之記憶體裝置，其中，當進行乘積累加運算時，對該編碼後輸入資料的各位元與該權重資料的該編碼後第一部份的各位元進行邏輯及運算；以及

當進行漢明距離運算時，對該編碼後輸入資料的各位元與該權重資料的該編碼後第一部份的各位元進行邏輯互斥或運算。

【請求項6】一種記憶體裝置之操作方法，包括：

第 2 頁，共 4 頁(發明申請專利範圍)

對一輸入資料進行編碼，將一編碼後輸入資料傳送至至少一頁緩衝器內，以及，從該至少一頁緩衝器平行讀出該編碼後輸入資料；

對一權重資料的一第一部份與一第二部份分別編碼為該權重資料的一編碼後第一部份與該權重資料的一編碼後第二部份，並將該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份寫入至該記憶體裝置之複數個記憶體晶胞內，以及，平行讀出該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份；

將該編碼後輸入資料分別乘上該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份，以平行產生複數個部份乘積；以及

將該些部份乘積累加，以產生一運算結果。

【請求項7】如請求項6所述之記憶體裝置之操作方法，其中，該權重資料的該第一部份是最高有效位元(most significant bit, MSB)，以及，該權重資料的該第二部份是最低有效位元(least significant bit, LSB)。

【請求項8】如請求項6所述之記憶體裝置之操作方法，其中，於編碼時，將該輸入資料與該權重資料分別量化成二進位整數向量；

將該輸入資料的各位元複製複數次並加上一備用位元；

將該權重資料分開成該第一部份與該第二部份；以及

將該權重資料的該第一部份與該第二部份的各位元以一元編碼表示，以得到該權重資料的該編碼後第一部份與該權重資料的該編碼後第二部份。

【請求項9】如請求項6所述之記憶體裝置之操作方法，其中，該運算結果包括一乘積累加結果

(multiply-and-accumulate，MAC)運算結果或者一漢明距離(Hamming distance)運算結果；以及

將屬同一維度的該些部份乘積累加，以得到單一維度乘積；

將複數個單一維度乘積累加，以得到一多維度乘積累加結果；

將複數個多維度乘積累加結果累加，以產生該運算結果。

【請求項10】如請求項9所述之記憶體裝置之操作方法，其中，當進行乘積累加運算時，對該編碼後輸入資料的各位元與該權重資料的該編碼後第一部份的各位元進行邏輯及運算；以及

當進行漢明距離運算時，對該編碼後輸入資料的各位元與該權重資料的該編碼後第一部份的各位元進行邏輯互斥或運算。

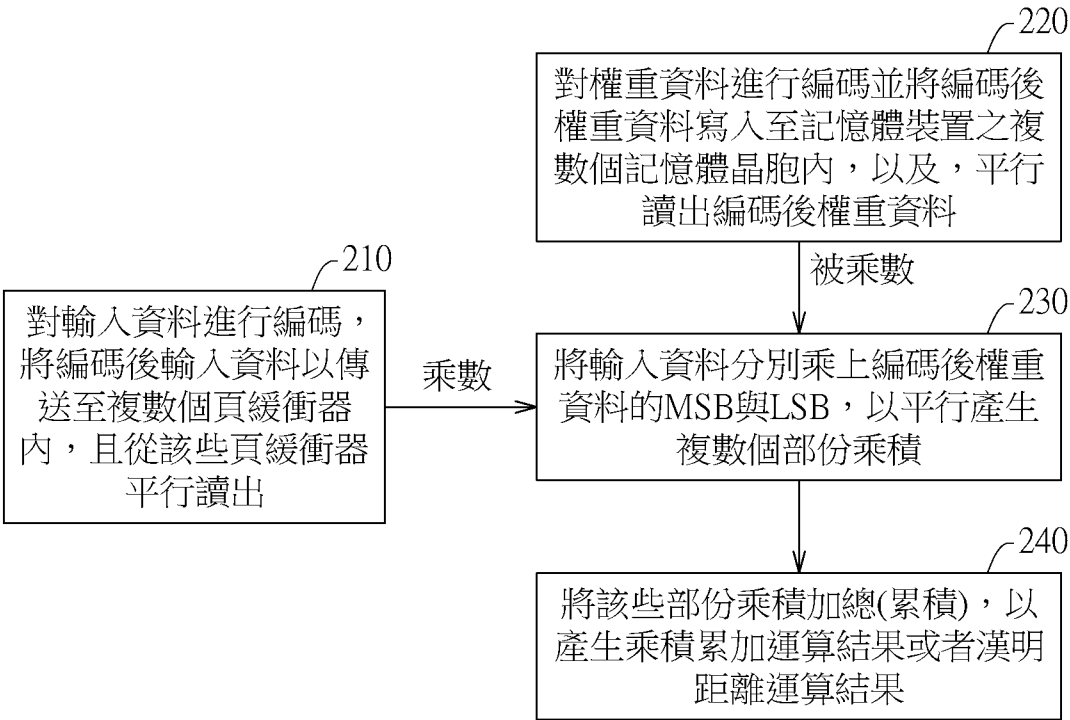
【發明圖式】

								p0[7]	p0[6]	p0[5]	p0[4]	p0[3]	p0[2]	p0[1]	p0[0]
								p1[7]	p1[6]	p1[5]	p1[4]	p1[3]	p1[2]	p1[1]	p1[0]
								+ p2[7]	p2[6]	p2[5]	p2[4]	p2[3]	p2[2]	p2[1]	p2[0]
								+ p3[7]	p3[6]	p3[5]	p3[4]	p3[3]	p3[2]	p3[1]	p3[0]
								+ p4[7]	p4[6]	p4[5]	p4[4]	p4[3]	p4[2]	p4[1]	p4[0]
								+ p5[7]	p5[6]	p5[5]	p5[4]	p5[3]	p5[2]	p5[1]	p5[0]
								+ p6[7]	p6[6]	p6[5]	p6[4]	p6[3]	p6[2]	p6[1]	p6[0]
								+ p7[7]	p7[6]	p7[5]	p7[4]	p7[3]	p7[2]	p7[1]	p7[0]
P[15]	P[14]	P[13]	P[12]	P[11]	P[10]	P[9]	P[8]	P[7]	P[6]	P[5]	P[4]	P[3]	P[2]	P[1]	P[0]

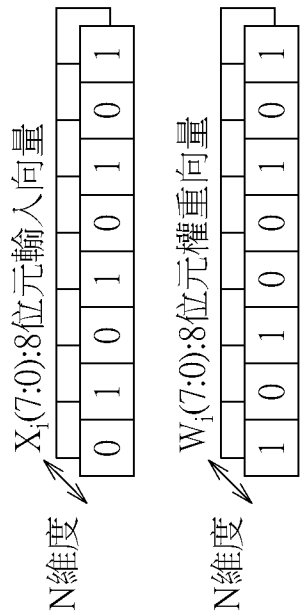
第1A圖

1	~p0[7] p0[6] p0[5] p0[4] p0[3] p0[2] p0[1] p0[0]														
	~p1[7] +p1[6] +p1[5] +p1[4] +p1[3] +p1[2] +p1[1] +p1[0]														
	~p2[7] +p2[6] +p2[5] +p2[4] +p2[3] +p2[2] +p2[1] +p2[0]														
	~p3[7] +p3[6] +p3[5] +p3[4] +p3[3] +p3[2] +p3[1] +p3[0]														
	~p4[7] +p4[6] +p4[5] +p4[4] +p4[3] +p4[2] +p4[1] +p4[0]														
	~p5[7] +p5[6] +p5[5] +p5[4] +p5[3] +p5[2] +p5[1] +p5[0]														
	~p6[7] +p6[6] +p6[5] +p6[4] +p6[3] +p6[2] +p6[1] +p6[0]														
	~p7[7] +p7[6] +p7[5] +p7[4] +p7[3] +p7[2] +p7[1] +p7[0]														
	P[15] P[14] P[13] P[12] P[11] P[10] P[9] P[8] P[7] P[6] P[5] P[4] P[3] P[2] P[1] P[0]														

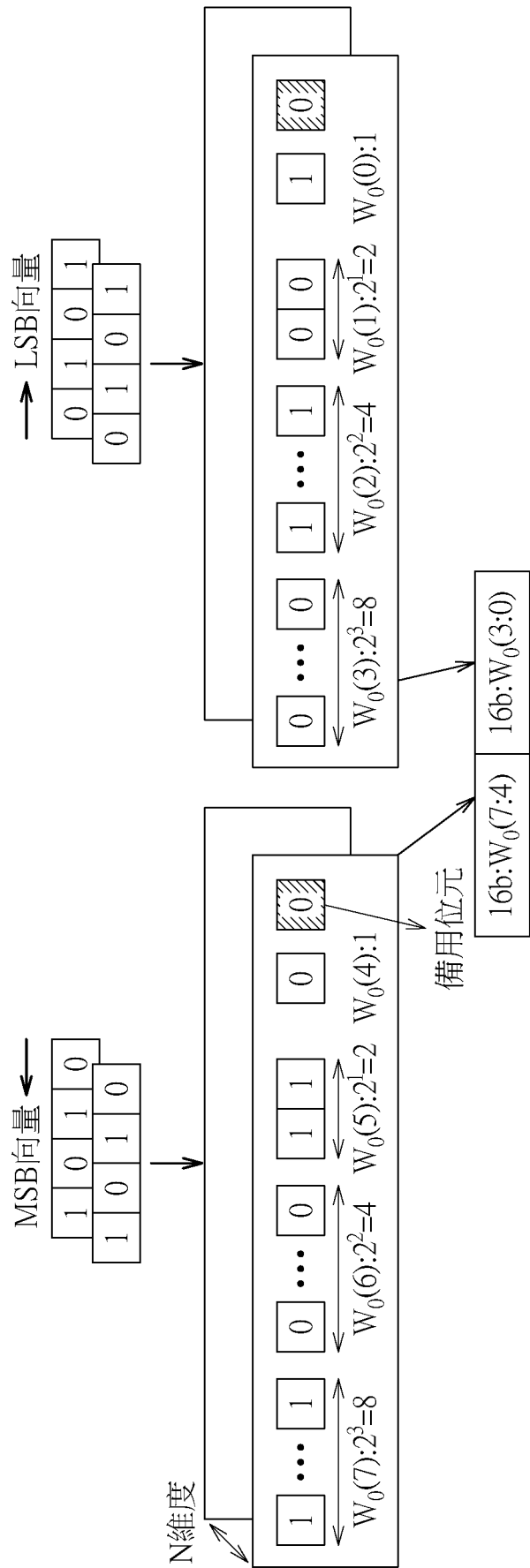
第 1B 圖



第 2 圖



第 3A 圖



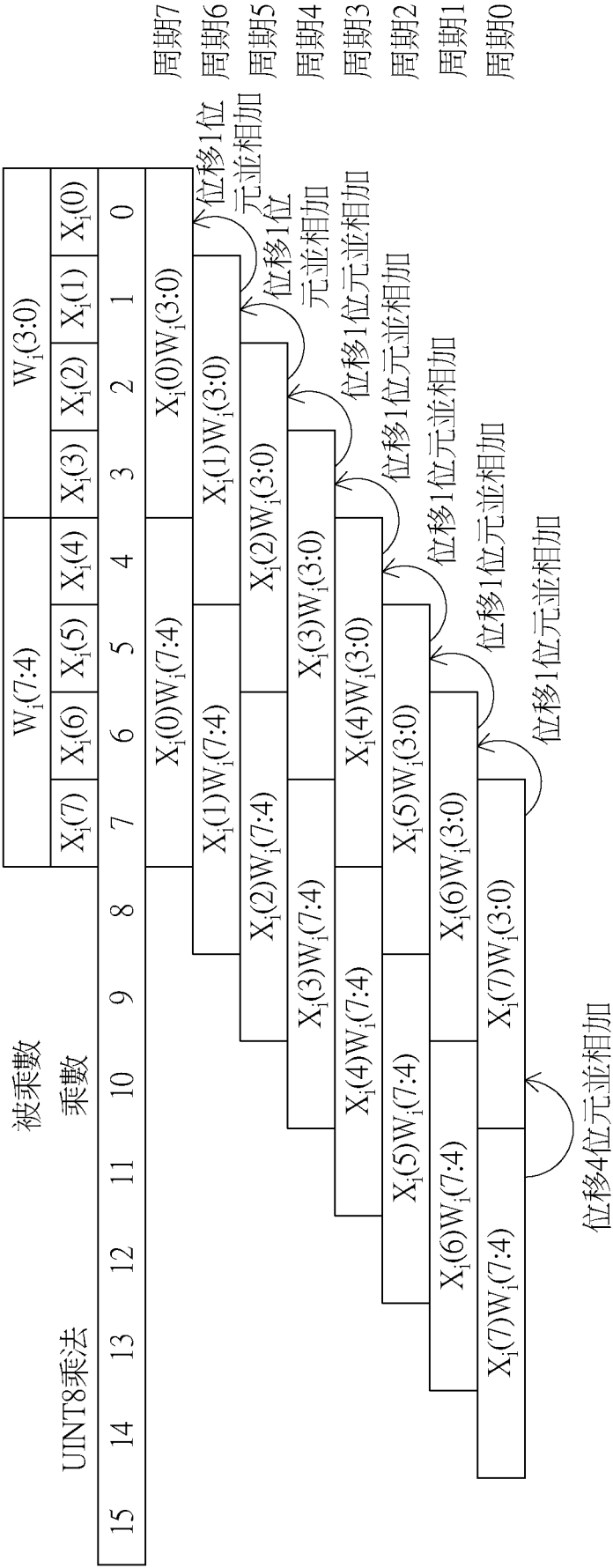
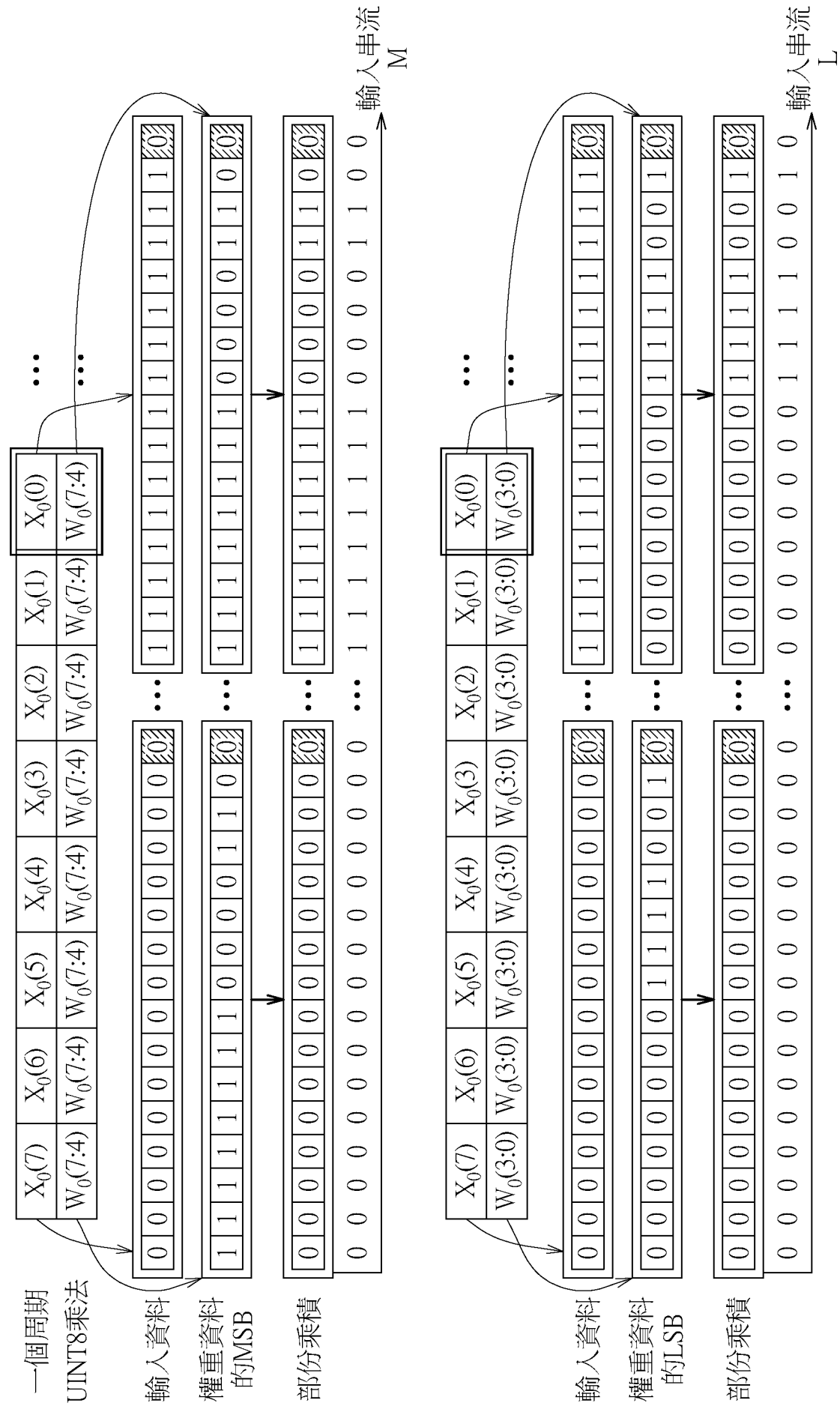
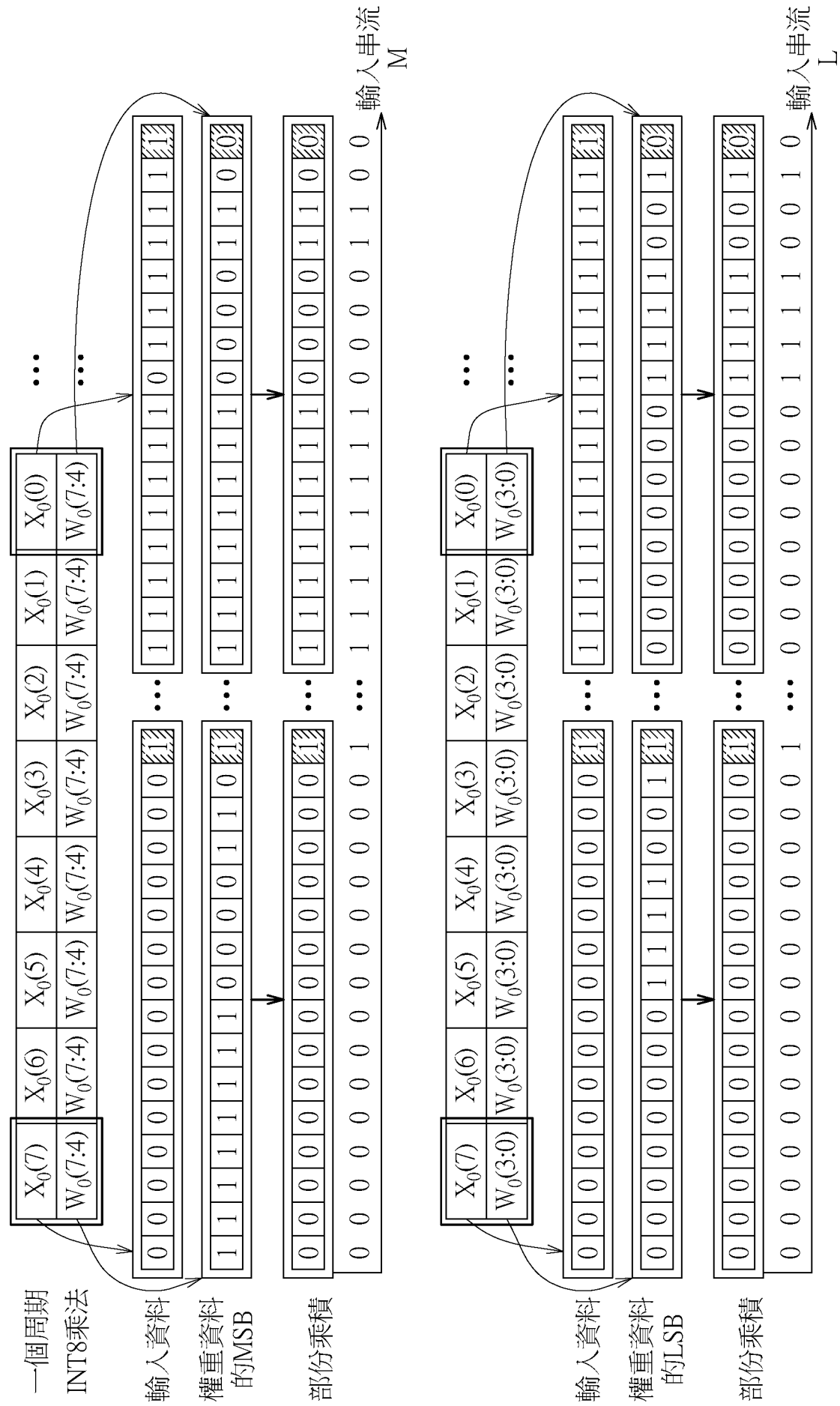


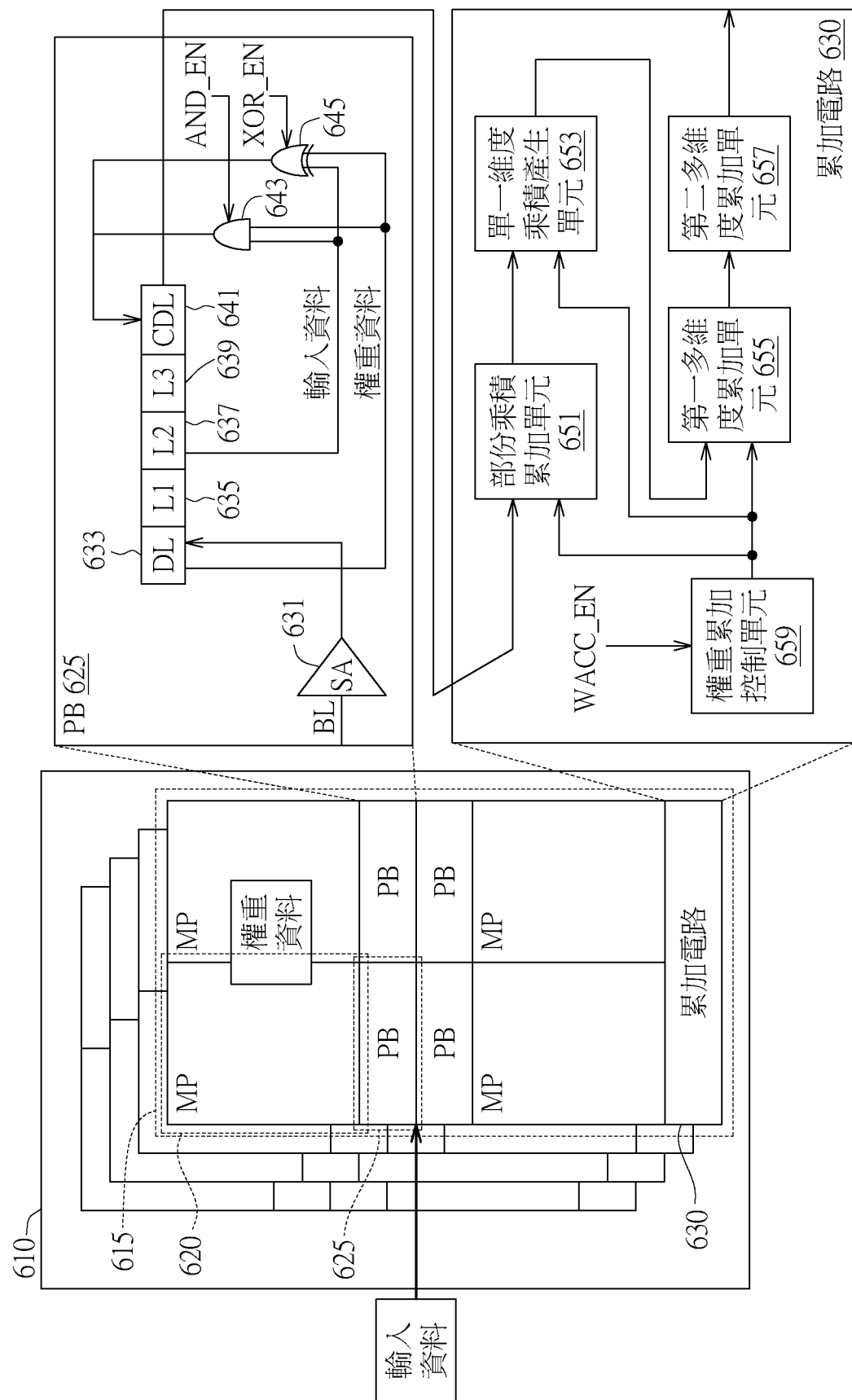
圖 4B 第

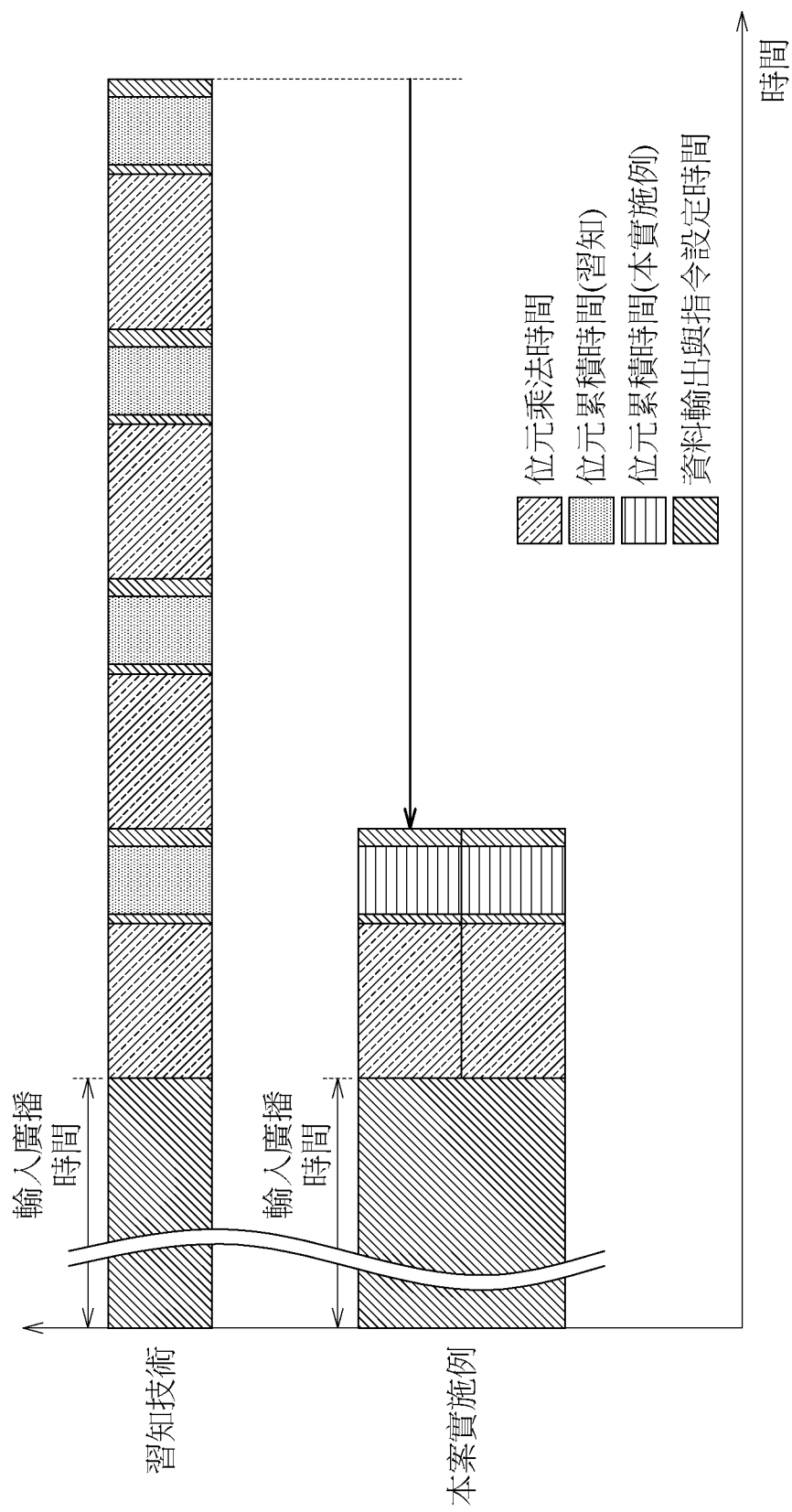


第5A圖

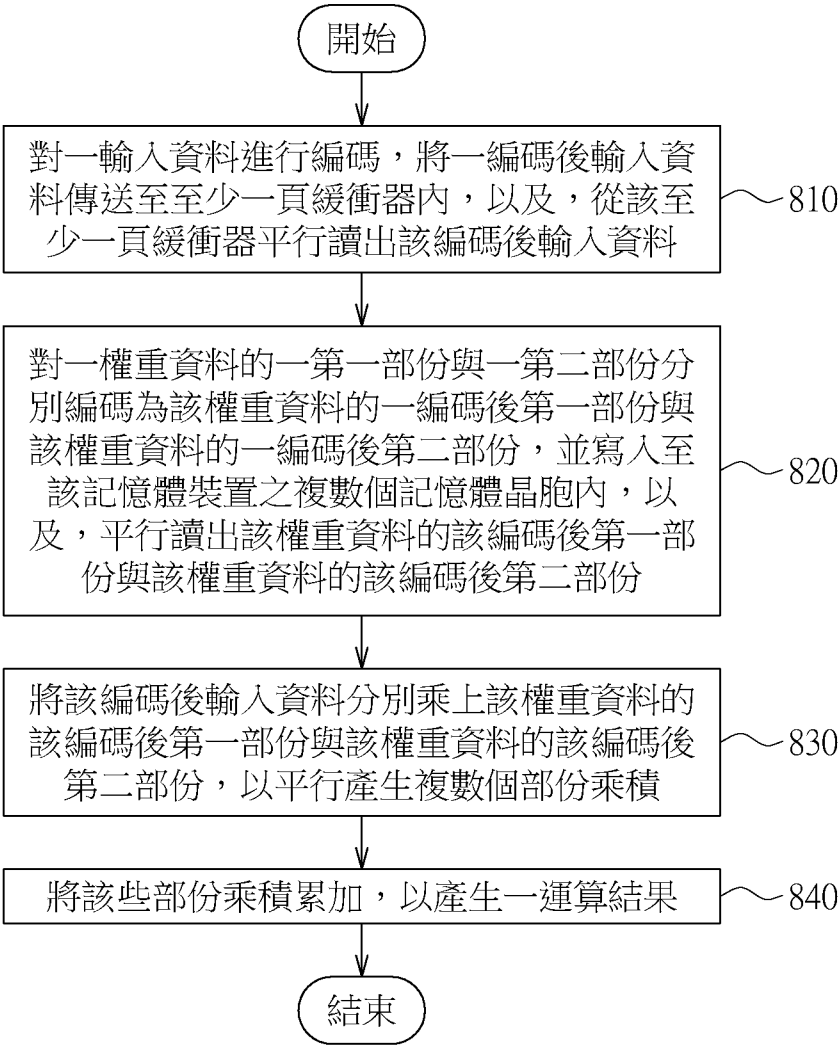


第5B圖

圖 6
採



第7圖



第 8 圖