



(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 I718336 B

(45)公告日：中華民國 110 (2021) 年 02 月 11 日

(21)申請案號：106130354

(22)申請日：中華民國 106 (2017) 年 09 月 06 日

(51)Int. Cl. : **G06F13/16 (2006.01)****G06F9/30 (2018.01)**

(30)優先權：2016/10/27 美國

62/413,973

2017/02/06 美國

15/426,015

(71)申請人：南韓商三星電子股份有限公司(南韓) SAMSUNG ELECTRONICS CO., LTD. (KR)
南韓(72)發明人：李双辰 LI, SHUANGCHEN (CN)；牛迪民 NIU, DIMIN (CN)；馬拉迪 克里希納
MALLADI, KRISHNA (IN)；郑宏忠 ZHENG, HONGZHONG (CN)

(74)代理人：林孟閱；盧佩君；陳怡如

(56)參考文獻：

TW 393605

TW 201532067A

TW 201633718A

US 5847577

US 6507530B1

US 2014/0208143A1

審查人員：吳兆平

申請專利範圍項數：20 項 圖式數：9 共 47 頁

(54)名稱

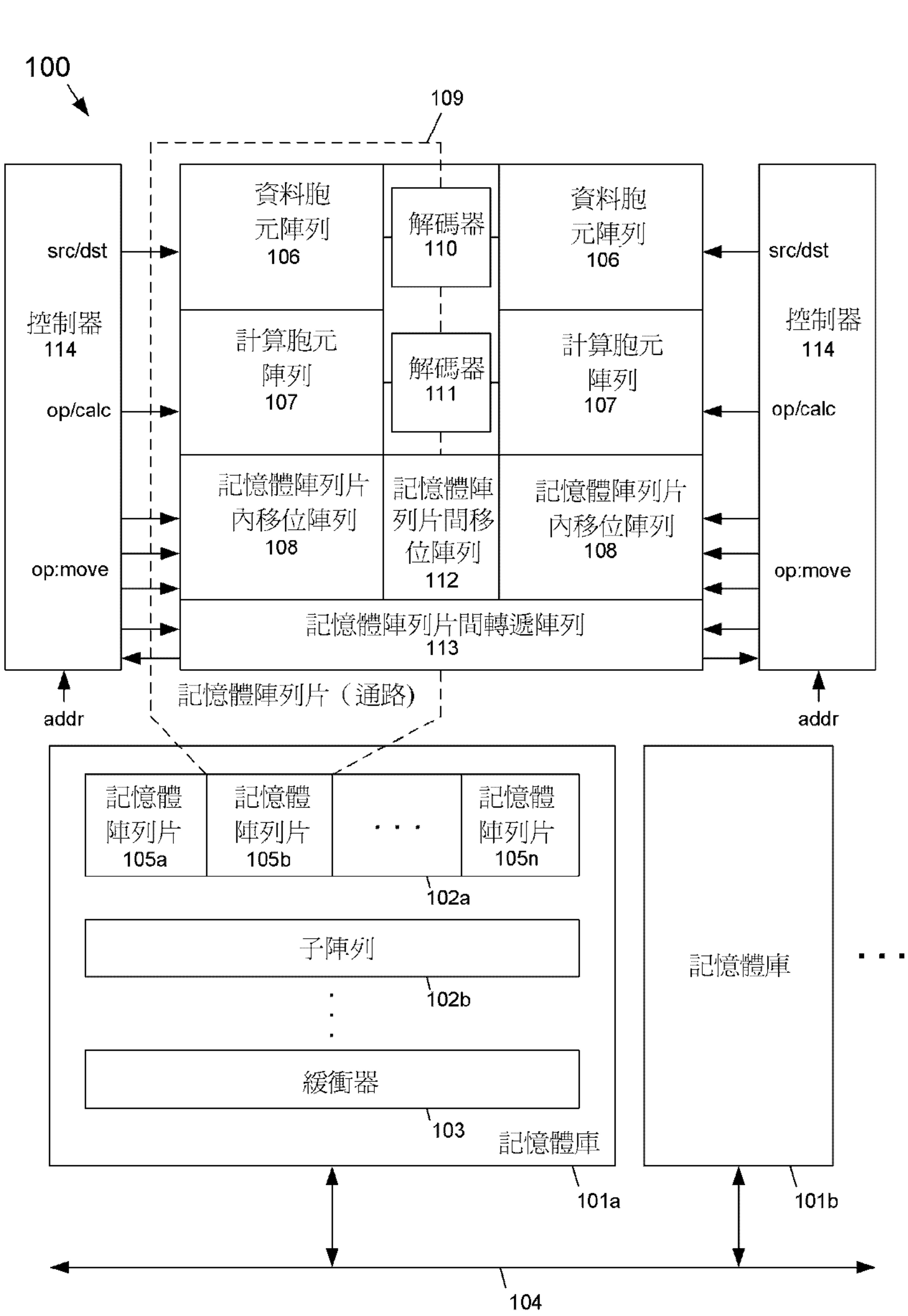
DPU 操作作用的系統

(57)摘要

一種系統，包含程式庫、編譯器、驅動器以及至少一個動態隨機存取記憶體(DRAM)處理單元(DPU)。程式庫可判定對應於所接收命令的至少一個 DPU 操作。編譯器可形成用於 DPU 操作的至少一個 DPU 指令。驅動器可發送至少一個 DPU 指令至至少一個 DPU。DPU 可包含至少一個計算胞元陣列，其包含以具有至少一個行的陣列配置的多個基於 DRAM 的計算胞元，其中至少一行可包含至少三個列的基於 DRAM 的計算胞元，至少三個列的基於 DRAM 的計算胞元經組態以提供在至少三個列的第一列及第二列上操作的邏輯功能且經組態以在至少三個列中的第三列中儲存邏輯功能的結果。

A system includes a library, a compiler, a driver and at least one dynamic random access memory (DRAM) processing unit (DPU). The library may determine at least one DPU operation corresponding to a received command. The compiler may form at least one DPU instruction for the DPU operation. The driver may send the at least one DPU instruction to at least one DPU. The DPU may include at least one computing cell array that includes a plurality of DRAM-based computing cells arranged in an array having at least one column in which the at least one column may include at least three rows of DRAM-based computing cells configured to provide a logic function that operates on a first row and a second row of the at least three rows and configured to store a result of the logic function in a third row of the at least three rows.

指定代表圖：



符號簡單說明：

100:DPU

101a、101b:記憶體庫

102a、102b:子陣列

103:緩衝器

104:系統匯流排

105a、105b、105n:記憶體陣列片

106:資料胞元陣列

107:計算胞元陣列

108:記憶體陣列片內移位陣列

109:虛線

110:資料胞元陣列解碼器

111:計算胞元陣列解碼器

112:記憶體陣列片間移位陣列

113:記憶體陣列片間轉遞陣列

114:子陣列控制器

addr:位址

【圖1】



I718336

【發明摘要】

【中文發明名稱】 DPU操作作用的系統

【英文發明名稱】 SYSTEM FOR DPU OPERATIONS

【中文】一種系統，包含程式庫、編譯器、驅動器以及至少一個動態隨機存取記憶體（DRAM）處理單元（DPU）。程式庫可判定對應於所接收命令的至少一個DPU操作。編譯器可形成用於DPU操作的至少一個DPU指令。驅動器可發送至少一個DPU指令至至少一個DPU。DPU可包含至少一個計算胞元陣列，其包含以具有至少一個行的陣列配置的多個基於DRAM的計算胞元，其中至少一行可包含至少三個列的基於DRAM的計算胞元，至少三個列的基於DRAM的計算胞元經組態以提供在至少三個列的第一列及第二列上操作的邏輯功能且經組態以在至少三個列中的第三列中儲存邏輯功能的結果。

【英文】 A system includes a library, a compiler, a driver and at least one dynamic random access memory (DRAM) processing unit (DPU). The library may determine at least one DPU operation corresponding to a received command. The compiler may form at least one DPU instruction for the DPU operation. The driver may send the at least one DPU instruction to at least one DPU. The DPU may include at least one computing cell array that includes a plurality of DRAM-based computing cells arranged in an array having at least one column in which the at least one column may include at least three rows of DRAM-based computing cells configured to provide a logic function that

operates on a first row and a second row of the at least three rows and configured to store a result of the logic function in a third row of the at least three rows.

【指定代表圖】圖1。

【代表圖之符號簡單說明】

100：DPU

101a、101b：記憶體庫

102a、102b：子陣列

103：緩衝器

104：系統匯流排

105a、105b、105n：記憶體陣列片

106：資料胞元陣列

107：計算胞元陣列

108：記憶體陣列片內移位陣列

109：虛線

110：資料胞元陣列解碼器

111：計算胞元陣列解碼器

112：記憶體陣列片間移位陣列

113：記憶體陣列片間轉遞陣列

114：子陣列控制器

addr：位址

【特徵化學式】

無

【發明說明書】

【中文發明名稱】 DPU操作作用的系統

【英文發明名稱】 SYSTEM FOR DPU OPERATIONS

[相關申請案的交叉參考]

【0001】 本專利申請案主張 2016 年 10 月 27 日申請的美國臨時專利申請案第 62/413,973 號的優先權權益，其全部揭露內容以引用的方式併入本文中。

【技術領域】

【0002】 本發明的實例實施例的態樣是關於一種用於整合及利用基於 DRAM 的處理單元（DRAM-Based Processing Unit; DPU）的軟體堆疊及程式化。

【先前技術】

【0003】 圖形處理單元（Graphics Processing Unit；GPU）及張量處理單元（Tensor Processing Unit；TPU）習知地用於深度學習處理。深度學習處理包含可能無法由 GPU 或 TPU 有效執行的高度並行化處理。

【發明內容】

【0004】 實例實施例提供一種可包含用以接收命令的介面、程式庫、編譯器、驅動器以及動態隨機存取記憶體（DRAM）處理單

元（DPU）的系統。程式庫可判定對應於由介面接收的命令的至少一個 DPU 操作。編譯器可形成用於對應於所接收命令的經判定至少一個 DPU 操作的至少一個 DPU 指令。驅動器可發送至少一個 DPU 指令至至少一個 DPU。DPU 可包含至少一個計算胞元陣列，所述至少一個計算胞元陣列包含以具有至少一個行的陣列方式配置的多個基於 DRAM 的計算胞元，其中所述至少一個行可包含至少三個列的基於 DRAM 的計算胞元，所述至少三個列的基於 DRAM 的計算胞元經組態以提供在至少三個列的第一列及第二列上操作的邏輯功能且經組態以在至少三個列中的第三列中儲存邏輯功能的結果。在一個實施例中，至少一個行的基於 DRAM 的計算胞元各自可包含三電晶體及一電容器（3T1C）式 DRAM 記憶體胞元，或一電晶體及一電容器（1T1C）式 DRAM 記憶體胞元，且至少一個行的基於 DRAM 的計算胞元可提供反或（NOR）邏輯功能。在一個實施例中，DPU 可更包括以具有至少一個行的陣列方式配置的多個基於 DRAM 的隨機計算胞元，其中所述至少一個行可包含至少三個列的基於 DRAM 的隨機計算胞元，所述至少三個列的基於 DRAM 的隨機計算胞元經組態以提供在至少三個列的第一列及第二列上操作的隨機邏輯功能且經組態以在至少三個列中的第三列中儲存隨機邏輯功能的結果。在一個實施例中，至少一個行的基於 DRAM 的隨機計算胞元各自可包含三電晶體及一電容器（3T1C）式 DRAM 記憶體胞元，或一電晶體及一電容器（1T1C）式 DRAM 記憶體胞元。在一個實施例中，至少一個 DPU 操作可包含隨機計算操作。

【0005】 實例實施例提供一種系統，可包括至少一個動態隨機存

取記憶體（DRAM）處理單元（DPU）、介面、程式庫、以及驅動器。每一 DPU 可包含至少一個計算胞元陣列。至少一個計算胞元陣列可包含：以具有至少一個行的陣列方式配置的多個基於 DRAM 的計算胞元，所述至少一個行可包含至少三個列的基於 DRAM 的計算胞元，所述至少三個列的基於 DRAM 的隨機計算胞元經組態以提供在所述至少三個列的第一列及第二列上操作的邏輯功能且經組態以在至少三個列中的第三列中儲存邏輯功能的結果；以及以具有至少一個行的陣列方式配置的多個基於 DRAM 的隨機計算胞元，其中至少一個行可包含至少三個列的基於 DRAM 的隨機計算胞元，所述至少三個列的基於 DRAM 的隨機計算胞元經組態以提供在至少三個列的第一列及第二列上操作的隨機邏輯功能且經組態以在至少三個列中的第三列中儲存隨機邏輯功能的結果。介面用以接收命令。程式庫用以判定對應於接收的命令的至少一個 DPU 操作。驅動器用以發送對應於所接收命令的至少一個 DPU 指令至至少一個 DPU。在一個實施例中，至少一個行的基於 DRAM 的計算胞元各自可包含三電晶體及一電容器（3T1C）式 DRAM 記憶體胞元，或一電晶體及一電容器（1T1C）式 DRAM 記憶體胞元，且至少一個行的基於 DRAM 的隨機計算胞元各自可包含三電晶體及一電容器（3T1C）式 DRAM 記憶體胞元，或一電晶體及一電容器（1T1C）式 DRAM 記憶體胞元。在一個實施例中，至少一個行的基於 DRAM 的計算胞元提供反或邏輯功能。在一個實施例中，至少一個 DPU 操作可包含隨機計算操作。

【0006】 實例實施例提供一種系統，可包含：至少一個動態隨機存取記憶體（DRAM）處理單元（DPU），所述 DPU 可包含至少一

個隨機計算胞元陣列，所述至少一個隨機計算胞元陣列包括以具有至少一個行的陣列方式配置的多個基於 DRAM 的隨機計算胞元，其中至少一個行可包含至少三個列的基於 DRAM 的隨機計算胞元，所述至少三個列的基於 DRAM 的隨機計算胞元經組態以提供在至少三個列的第一列及第二列上操作的隨機邏輯功能且經組態以在至少三個列中的第三列中儲存隨機邏輯功能的結果；程式庫，用以判定對應於接收的命令的至少一個隨機 DPU 操作；以及驅動器，用以發送對應於接收的命令的至少一個隨機 DPU 指令至包括多個基於 DRAM 的隨機計算胞元的至少一個 DPU。在一個實施例中，至少一個行的基於 DRAM 的隨機計算胞元各自可包含三電晶體及一電容器（3T1C）式 DRAM 記憶體胞元，或一電晶體及一電容器（1T1C）式 DRAM 記憶體胞元。

【圖式簡單說明】

【0007】 在以下部分中，將參考圖式中所說明的例示性實施例描述本文中所揭露的主題的態樣，其中：

【0008】

圖 1 描繪根據本文中所揭露主題的基於動態隨機存取記憶體（DRAM）的處理單元（DPU）的實例實施例的方塊圖。

圖 2A 描繪可以用於計算胞元陣列中的計算胞元的三電晶體一電容器 DRAM 計算胞元構形的實例實施例。

圖 2B 描繪可以用於計算胞元陣列中的計算胞元的一電晶體一電容器 DRAM 計算胞元構形的替代實例實施例。

圖 3 描繪根據本文中所揭露主題的記憶體陣列片內移位陣列

的實例實施例。

圖 4A 描繪根據本文中所揭露主題的記憶體陣列片間移位陣列的實施例。

圖 4B 概念地描繪根據本文中所揭露主題的用於左記憶體陣列片間移位的相鄰計算胞元行中的兩個相同定位計算胞元之間的記憶體陣列片間移位互連組態。

圖 4C 概念地描繪根據本文中所揭露主題的在用於左記憶體陣列片間移位的相鄰計算胞元行中的兩個非相同定位計算胞元之間的記憶體陣列片間移位互連組態。

圖 5 描繪根據本文中所揭露主題的記憶體陣列片間轉遞陣列的實施例。

圖 6A 至圖 6G 描繪根據本文中所揭露主題的可由 DPU 提供的基於反或邏輯的操作。

圖 7 描繪根據本文中所揭露主題的包括隨機資料區的 DPU 的實例實施例的方塊圖。

圖 8A 及圖 8B 分別描繪用於可轉換成多工操作的加法操作及可轉換成及（AND）邏輯操作的乘法操作的隨機計算操作。

圖 9 描繪根據本文中所揭露主題的包括 DPU 的系統架構。

【實施方式】

【0009】 在以下詳細描述中，闡述眾多特定細節以便提供對本發明的透徹理解。然而，熟習此項技術者將理解，可在無此等特定細節的情況下實踐所揭露態樣。在其他情況下，為了不混淆本文中所揭露的主題，尚未詳細描述熟知方法、程序、組件以及電路。

【0010】貫穿本說明書對「一個實施例」或「一實施例」的參考意謂結合實施例所描述的特定特徵、結構或特性可包括於本文中所揭露的至少一個實施例中。因此，片語「在一個實施例中」或「在一實施例中」或「根據一個實施例」（或具有類似意思的其他片語）貫穿本說明書在各種位置的出現可能未必全部指同一實施例。另外，特定特徵、結構或特性可在一或多個實施例中以任何合適的方式組合。就此而言，如本文中所使用，詞語「例示性」意謂「充當實例、例子或說明」。不應將本文中描述為「例示性」的任一實施例解釋為有必要比其他實施例較佳或有利。另外，取決於本文中論述的上下文，單數術語可包含對應複數形式，且複數術語可包含對應單數形式。另外應注意，本文中所展示及論述的各種圖式（包含組件圖）僅用於說明性目的，且未按比例繪製。類似地，僅用於說明性目的而展示各種波形及時序圖。舉例而言，為了清楚起見，可相對於其他元件放大一些元件的尺寸。另外，在認為適當時，已在諸圖中重複參考編號以指示對應及/或類似元件。

【0011】本文中所使用的術語僅出於描述特定例示性實施例的目的，且不意欲限制所主張主題。如本文中所使用，除非上下文另有清晰地指示，否則單數形式「一」以及「所述」意欲亦包含複數形式。應進一步理解，術語「包括」在用於本說明書中時規定所陳述的特徵、整數、步驟、操作、元件及/或組件的存在，但不排除一個或多個其他特徵、整數、步驟、操作、元件、組件及/或其群組的存在或添加。如本文中所使用，術語「第一」、「第二」等被用作所述術語在前的名詞的標記，且不暗示任何類型的排序

(例如，空間、時間、邏輯等)，除非明確地如此定義。此外，相同參考數字可用於跨越兩個或多於兩個圖指具有相同或類似功能性的部分、組件、區塊、電路、單元或模組。然而，此用法僅為簡化說明且易於論述起見；其不暗示此類組件或單元的建構或建築結構細節跨越所有實施例而相同，或此類通常被參考的部分/模組為實施本文中所揭露的特定實施例的教示的唯一方式。

【0012】 除非另有定義，否則本文中使用的所有術語（包括技術及科學術語）具有與由熟習此主題所屬領域的一般技術者通常所理解相同的意義。應進一步理解，諸如常用詞典中所定義的術語的術語應被解釋為具有與其在相關技術的上下文中的意義一致的意義，且將不在理想化或過度正式意義上進行解釋，除非本文中明確地如此界定。

【0013】 本文中所揭露主題提供可程式化及可重組態用於不同操作（諸如但不限於加法、乘法、移位、最大/最小，以及比較）的基於動態隨機存取記憶體（**DRAM**）的處理單元（**DPU**）。在一個實施例中，**DPU** 是基於三電晶體及一電容器（**3T1C**）式 **DRAM** 程序及結構。在另一實施例中，**DPU** 是基於具有微幅修改的一電晶體及一電容器（**1T1C**）式 **DRAM** 程序及結構。因此，**DPU** 不含有特定計算邏輯電路（如加法器），但提供使用使用高度並行操作的記憶體胞元的計算。在一個實施例中，**DPU** 可包含其中加法可轉換成多工操作且乘法可轉換成及邏輯操作的隨機計算陣列。

【0014】 本文中所揭露主題亦提供系統架構，其包含具有框架延伸的環境（生態系統）、程式庫、驅動器、編譯器以及指令集架構（**ISA**）以程式化並重新組態 **DPU**。

【0015】 另外，本文中所揭露主題提供一種系統架構，其適合於資料中心及/或行動應用且提供替代 GPU/ASIC (TPU) /FPGA 機器學習應用的用於二進位及固定點計算的機器學習應用的記憶體中處理器 (Processor-in-Memory ; PIM) 方案。在一個實施例中，本文中所揭露主題提供高效能、能效及低成本系統，其提供用於例如二進位權重神經網路的加速深度學習。

【0016】 本文中所揭露主題是關於可使用動態隨機存取記憶體 (DRAM) 技術形成且可重組態及可程式化的基於 DRAM 的處理單元 (DPU)。在一個實施例中，DPU 可包含基於 DRAM 的記憶體胞元陣列及可經組態以執行不同操作 (諸如加法、乘法、分類等) 的基於 DRAM 的計算胞元陣列。

【0017】 DPU 的內部架構可包含連接至多個子陣列記憶體庫的系統匯流排。在一個實施例中，系統匯流排可經組態以提供子陣列的 H 樹連接記憶體庫。每一子陣列可包含本地控制器，且每一個別子陣列可單獨地或同時被啟動。在一個實施例中，基於 DRAM 的胞元可分成兩個陣列，資料胞元陣列及計算胞元陣列。在一個實施例中，計算胞元陣列可藉由基於 DRAM 的記憶體胞元實施。在另一實施例中，計算胞元陣列可藉由具有邏輯電路的基於 DRAM 的記憶體胞元實施。DPU 內部架構亦可包含資料移位及資料移動電路。在一些實施例中，可存在可經組態用於隨機資料計算的第三基於 DRAM 的胞元陣列。

【0018】 圖 1 描繪根據本文中所揭露主題的 DPU 100 的實例實施例的方塊圖。DPU 100 可包含一或多個記憶體庫 101a 至 101m，其中在圖 1 中僅僅描繪記憶體庫 101a 及 101b。每一記憶體庫 101

可包含一或多個子陣列 102a 至 102n，其中在圖 1 中僅僅描繪子陣列 102a 及子陣列 102b。每一記憶體庫 101 亦可包含緩衝器 103。緩衝器 103 可耦接至個別子陣列 102 及系統匯流排 104。緩衝器 103 可讀取記憶體庫 102 中的整個列，且接著將列寫回至同一記憶體庫或另一記憶體庫。緩衝器 103 亦可廣播列資料的複本至子陣列 102 中的多個記憶體陣列片 105a 至 105n。在一個實施例中，記憶體庫 101 及系統匯流排 104 可經組態以提供 H 樹連接記憶體庫。

【0019】 每一子陣列 102 可包含一或多個記憶體陣列片(或通路) 105，其中在圖 1 中描繪子陣列 102a 的記憶體陣列片 105a 至 105n。每一記憶體陣列片 105 為可包含資料胞元陣列 106、計算胞元陣列 107 以及記憶體陣列片內移位陣列 108 的 DPU 100 的區。實例記憶體陣列片 105 在圖 1 中被指示為由虛線 109 圍封。每一記憶體陣列片 105 可與相鄰記憶體陣列片共用資料胞元陣列解碼器 110、計算胞元陣列解碼器 111、記憶體陣列片間移位陣列 112 以及記憶體陣列片間轉遞陣列 113。在一個實施例中，資料胞元陣列解碼器 110、計算胞元陣列解碼器 111 以及記憶體陣列片間移位陣列 112 可與子陣列控制器 114 一起交替實體地配置於相鄰記憶體陣列片 105 之間。在一個實施例中，解碼器 110 及解碼器 111 可操作為習知 DRAM 型記憶體解碼器。

【0020】 在一個實施例中，每一記憶體陣列片 105 通信地耦接至子陣列控制器 114。每一子陣列控制器 114 可經組態為獨立於其他子陣列控制器 114。子陣列控制器 114 可自 DRAM 位址匯流排接收作為位址 (addr) 的指令。回應於位址 (亦即，位址信號)，子

陣列控制器 114 可將經解碼位址作為輸出提供至資料胞元陣列 106 及計算胞元陣列 107 中的任一者或兩者。亦即，子陣列控制器 114 可輸出藉由用於相關聯資料胞元陣列 106 的解碼器 110 解碼的來源/目的地（src/dst）位址，且在計算胞元陣列 107 的情況下，可輸出藉由解碼器 111 解碼的操作/計算（op/cal）位址。子陣列控制器 114 亦可自 DRAM 匯流排接收作為位址的指令，其使兩個或大於兩個子陣列控制器 114 以協調方式操作。子陣列控制器 114 亦可控制資料移動電路，諸如控制記憶體陣列片內移位陣列 108、記憶體陣列片間移位陣列 112 以及記憶體陣列片間轉遞陣列 113。

【0021】 每一資料胞元陣列 106 可包含在至少一個行及至少一個列中配置的一或多個動態隨機存取記憶體（DRAM）胞元。在一個實施例中，資料胞元陣列 106 可經組態為習知 DRAM 胞元陣列。在一個實施例中，資料胞元陣列 106 可包含 2K 行及 16 列。在另一實施例中，資料胞元陣列 106 可包含少於或多於 2K 行及/或少於或多於 16 列。

【0022】 每一計算胞元陣列 107 可包含在至少一個行及至少一個列中配置的一或多個計算胞元。計算胞元陣列 107 中行的數目與資料胞元陣列 106 中行的數目相同。在一個實施例中，計算胞元陣列 107 可包含 2K 行及 16 列。在另一實施例中，計算胞元陣列 107 可包含少於或多於 2K 行及/或少於或多於 16 列。

【0023】 圖 2A 描繪可以用於計算胞元陣列 107 中的計算胞元的三電晶體及一電容器（3T1C）式 DRAM 計算胞元構形 201 的實例實施例。如圖 2A 中所描繪，在列 X 中的 3T1C 計算胞元包含具有電耦接至寫入位元線（寫入 BL（Write BL））的源極端子、電耦接至

電容器 C_1 的第一端子及第二電晶體 T_2 的閘極端子的汲極端子以及電耦接至寫入啟用線 (WEN_x) 的閘極端子的第一電晶體 T_1 。電容器 C_1 的第二端子電耦接至接地線。電晶體 T_2 包含電耦接至接地線的源極端子，以及電耦接至第三電晶體 T_3 的源極端子的汲極端子。第三電晶體 T_3 包含電耦接至字線 (WL_x) 的閘極端子，以及電耦接至讀取位元線 (讀取 BL ($Read\ BL$)) 的汲極端子。3T1C 計算胞元構形 201 包含具有電耦接至讀取 BL 的輸入端及電耦接至寫入 BL 的輸出端的感測放大器 SA 。

【0024】 在列 Y 中的計算胞元及在列 R 中的計算胞元亦可包含類似於列 X 中的計算胞元的配置以 3T1C DRAM 組態配置的三個電晶體 T_1 至 T_3 及電容器 C_1 。圖 2A 中描繪的實例三個計算胞元及感測放大器 SA 經組態以提供反或邏輯操作，亦即 X 反或 Y 邏輯操作，其中結果儲存於列 R 中。儘管 3T1C DRAM 計算胞元的僅僅一行經明確地描繪於圖 2A 中，但應理解在另一實施例中，3T1C 計算胞元可經組態成多個行 (亦即， $2K$ 行)。亦應理解在另一實施例中，可提供大於三列。另外，雖然圖 2A 中描繪的 3T1C DRAM 計算胞元組態提供反或邏輯操作，但應理解 3T1C DRAM 計算胞元構形 201 的反或邏輯操作可用以提供功能操作，諸如但不限於互斥反或 ($XNOR$)、加法 (ADD)、選擇 (SET)、最大值 (MAX)、正負號 ($SIGN$)、多工 (MUX)、條件和加法邏輯 (CSA)、乘法、位 1 計數以及比較 ($COMPARE$)。移位陣列 108 及 112 亦提供移位功能。

【0025】 圖 2B 描繪可以用於圖 1 的計算胞元陣列 107 中的計算胞元的一電晶體及一電容器 (1T1C) 式 DRAM 計算胞元構形 202 的

替代實例實施例。如圖 2B 中所描繪，1T1C DRAM 計算胞元包含具有電連接至電容器 C_2 的第一端的源極端子、電連接至位元線（BL）的汲極端子以及電連接至字線（WL）的閘極端子的電晶體 T_4 。電容器 C_2 的第二端子電耦接至接地線。位元線 BL 電耦接至感測放大器 SA 的輸入端。感測放大器 SA 的輸出端電耦接至多工器（MUX）的第一輸入端、電晶體 T_5 的汲極端子以及算術邏輯單元（ALU）的輸入端。MUX 的輸出端電耦接至栓鎖器（LATCH）的輸入端。電晶體 T_5 的源極端子電耦接至栓鎖器的輸出端。ALU 的輸出端電耦接至 MUX 的第二輸入端。圖 2B 中的電晶體 T_5 、MUX、LATCH 及 ALU 各自分別自控制器 114 接收控制信號 CNTL1 至 CNTL4。在一個實施例中，ALU 可經組態以提供反或功能。儘管電耦接至圖 2B 中的位元線 BL 的邏輯電路提供反或邏輯操作，但應理解電耦接至位元線 BL 的邏輯電路（亦即 ALU）可提供其他功能操作，諸如但不限於互斥反或（XNOR）、加法（ADD）、選擇（SET）、MAX、SIGN、多工（MUX）、條件和加法邏輯（CSA）、乘法、位 1 計數以及 COMPARE。移位陣列 108 及 112 亦提供移位功能。應理解僅僅一個 1T1C 計算胞元在圖 2B 中描繪且可提供 1T1C 計算胞元的多個行及列。

【0026】 如可在圖 2A 及圖 2B 中看到，DPU 的計算胞元不包含特定複雜計算邏輯，但實際上包含具有提供執行多個不同類型計算能力的可重程式化性質的相對簡單構形。另外，DPU 構形可經配置以利用記憶體結構中固有的大規模並行度以更快且更有效地執行更多計算。

【0027】 圖 3 描繪根據本文中所揭露主題的記憶體陣列片內移位

陣列 108 實例實施例。為了簡化記憶體陣列片內移位陣列 108 的描述，考慮為四行計算胞元 107 寬的記憶體陣列片 105，諸如圖 3 中描繪。記憶體陣列片內移位陣列 108 包含以陣列方式配置的多個電晶體 T_6 (其中圖 3 中指示僅僅一個電晶體 T_6)， $2n$ 移位線 SLs (其中 n 為記憶體陣列片 105 中的計算胞元的行)， $n+2$ 個左移位控制線 $SLcL$ ，2 個右移位控制線 $SRcL$ 以及 n 個移位遮罩線 SML 。記憶體陣列片內移位陣列 108 的電晶體 T_6 中的一些電連接於寫入 BL 與 $2n$ 移位線 SL 之間，且記憶體陣列片內移位陣列 108 的其他電晶體 T_6 連接於讀取 BL 與 $2n$ 移位線 SL 之間。此等電晶體 T_6 的閘極電耦接至 $n+2$ 個左移位控制線 $SLcL$ 及 2 個右移位控制線 $SRcL$ 。記憶體陣列片內移位陣列 108 的其他電晶體 T_6 電連接於 n 個移位遮罩線 SML 與 $2n$ 個移位線 SL 之間。記憶體陣列片內移位陣列 108 的控制線電耦接至與記憶體陣列片 105 相關聯的子陣列控制器 114。

【0028】 記憶體陣列片內移位陣列 108 可藉由控制線 $SLcL$ 及 $SRcL$ 上的適當信號在記憶體陣列片 105 內向左或向右移位資料。對於左移位，資料可以正負號位元填充，且每一操作移位 1 位元或 $(n-1)$ 位元，其中 n 為每一記憶體陣列片 105 的行的數目。對於右移位，資料可以如由指令控制的 0 或 1 填充，且每一記憶體陣列片 (MAT) 移位 2^0 、 2^1 、 $\dots$ 、 2^{k-1} 、 2^k 直至行的數目，其中 2^k 為行的數目。

【0029】 圖 4A 描繪根據本文中所揭露主題的記憶體陣列片間移位陣列 112 的實施例。為了簡化記憶體陣列片間移位陣列 112 的描述，考慮其中記憶體陣列片 105 寬度為兩行計算胞元 107 的組

態，諸如圖 4A 至圖 4C 中所描繪。亦即，每一記憶體陣列片 105 包含計算胞元 107a 的第一行及計算胞元 107b 的第二行。記憶體陣列片間移位陣列 112 包含電晶體 T_{112a} 及 T_{112b} 、電晶體 T_{112c} 及 T_{112d} 、資料移位線 112e 及 112f，以及記憶體陣列片間移位控制線 ISLcL。在記憶體陣列片內，電晶體 T_{112a} 包含電耦接至計算胞元 107a 的第一行的讀取 BL 的源極端子，電耦接至資料移位線 112e 的汲極端子。電晶體 T_{112b} 包含電耦接至計算胞元 107b 的第二行的讀取 BL 的源極端子，電耦接至資料移位線 112f 的汲極端子。資料移位線 112e 及 112f 電耦接至緩衝器 103(圖 4A 中未展示)。在不同記憶體陣列片之間，電晶體 T_{112c} 包含分別電耦接至相鄰記憶體陣列片中的資料移位線 112e 的源極端子及汲極端子。電晶體 T_{112d} 包含分別電耦接至相鄰記憶體陣列片中的資料移位線 112f 的源極端子及汲極端子。電晶體 T_{112c} 及 T_{112d} 的閘極分別電耦接至分別不同記憶體陣列片間移位控制線 ISLcL。記憶體陣列片間移位陣列 112 可藉由控制線 ISLcL 上的適當信號在不同記憶體陣列片之間向左或向右移位資料。記憶體陣列片間移位陣列 112 的控制線電耦接至與記憶體陣列片 105 相關聯的子陣列控制器 114。

【0030】 圖 4B 概念地描繪根據本文中所揭露主題的用於左記憶體陣列片間移位的相鄰計算胞元行 105a 及 105b 中的兩個相同定位計算胞元之間的記憶體陣列片間移位互連組態。圖 4B 的互連組態可藉由被加重的操作性互連節點概念地描繪。舉例而言，電晶體 T_{112c} 及 T_{112d} 被啟動以使得導電路徑存在於每一電晶體之間，藉此連接計算胞元行 105a（在左側）與 105b（在右側）之間的資料移位線 112e 及 112f。電晶體 T_{112c} 及 T_{112d} 的閘極端子電連接至作

用中記憶體陣列片間移位控制線 ISLcL。記憶體陣列片 105b 中的電晶體 T_{112a} 及 T_{112b} 被啟動以使得記憶體陣列片 105b 中的計算胞元 107a 的讀取 BL 電連接至記憶體陣列片 105b 左邊的記憶體陣列片 105a 中的計算胞元 107a 的寫入 BL，且使得記憶體陣列片 105b 中的計算胞元 107b 的讀取 BL 電連接至記憶體陣列片 105b 左邊的記憶體陣列片 105a 中的計算胞元 107b 的寫入 BL。

【0031】 圖 4C 概念地描繪根據本文中所揭露主題的在用於左記憶體陣列片間移位的相鄰計算胞元行 105a 及 105b 中的兩個非相同定位計算胞元之間的記憶體陣列片間移位互連組態。圖 4C 的互連組態可藉由被加重的操作性互連節點概念地描繪。舉例而言，電晶體 T_{112c} 及 T_{112d} 被啟動以使得導電路徑存在於每一電晶體之間，藉此連接計算胞元行 105a（在右側）與 105b（在左側）之間的資料移位線 112e 及 112f。電晶體 T_{112c} 及 T_{112d} 的閘極端子電連接至作用中記憶體陣列片間移位控制線 ISLcL。記憶體陣列片 105b 中的電晶體 T_{112a} 及 T_{112b} 被啟動以使得記憶體陣列片 105a 中的計算胞元 107a 的讀取 BL 電連接至記憶體陣列片 105a 左邊的記憶體陣列片 105b 中的計算胞元 107a 的寫入 BL，且使得記憶體陣列片 105a 中的計算胞元 107b 的讀取 BL 電連接至記憶體陣列片 105a 左邊的記憶體陣列片 105b 中的計算胞元 107a 的寫入 BL。

【0032】 圖 5 描繪根據本文中所揭露主題的記憶體陣列片間轉遞陣列 113 的實施例。為了簡化記憶體陣列片間轉遞陣列 113 的描述，考慮其中記憶體陣列片 105 為兩行計算胞元 107 寬的組態，諸如圖 5 中所描繪。亦即，每一記憶體陣列片 105 包含計算胞元 107a 的第一行及計算胞元 107b 的第二行。在記憶體陣列片 105 情

況下，記憶體陣列片間轉遞陣列 113 包括電晶體 T_{113a} 及 T_{113b} 、電晶體 T_{113c} 及 T_{113d} 以及電晶體 T_{113e} 及 T_{113f} 、 2^n 個資料轉遞線 FDL（其中 n 為記憶體陣列片中的計算胞元行的數目）、轉遞控制線 FCL，以及 2^m 個轉遞區段線 FSL（其中 m 為區段的數目）。電晶體 T_{113a} 及 T_{113b} 的源極端子分別電連接計算胞元 107a 的第一行的寫入 BL 及讀取 BL。電晶體 T_{113a} 及 T_{113b} 的汲極端子電耦接至第一資料轉遞線 FDL 113g。電晶體 T_{113c} 及 T_{113d} 的源極端子分別電連接計算胞元 107b 的第二行的寫入 BL 及讀取 BL。電晶體 T_{113c} 及 T_{113d} 的汲極端子電耦接至第二資料轉遞線 FDL 113h。電晶體 T_{113e} 及 T_{113f} 的源極端子分別電耦接至電晶體 T_{113a} 及 T_{113b} 的閘極端子。電晶體 T_{113e} 及 T_{113f} 的汲極端子兩者均耦接至相同轉遞第二線 FSL。電晶體 T_{113e} 及 T_{113f} 的閘極端子分別耦接至不同轉遞控制線 FCL。記憶體陣列片間轉遞陣列 113 可藉由轉遞控制線 FCL 上的適當信號在記憶體陣列片之間轉遞資料。記憶體陣列片間轉遞陣列 113 的控制線電耦接至與資料在其間轉遞的記憶體陣列片 105 相關聯的子陣列控制器 114。

【0033】 圖 6A 至圖 6G 描繪根據本文中所揭露主題的可由 DPU 提供的基於反或邏輯的操作。在圖 6A 至圖 6G 中，第一運算元可儲存於列 X 中且第二運算元可儲存於列 Y 或列 W 中。圖 6A 至圖 6G 中的箭頭表示計算胞元的整個列的反或邏輯操作的輸入及輸出流。舉例而言，圖 6A 中的列 X 可表示儲存於列 X 的計算胞元中的運算元的整個列。對儲存於列 X 中的運算元及儲存於列 Y 中的運算元的反或邏輯操作的結果儲存於結果列 R 中。在一個實施例中，列 X 及列 Y 中的運算元可包含例如 100 行（亦即， x_1 、 x_2 、 $\dots$ 、

x_{100} 及 $y_1、y_2、\cdots、y_{100}$) 且結果可儲存於列 R (亦即, $r_1、r_2、\cdots、r_{100}$) 中。亦即, $x_i \text{ NOR } y_i = r_i$, 其中 i 為行索引。在另一實施例中, 列 X 可表示列中的計算胞元的僅僅選定群組。

【0034】 圖 6B 描繪基於前綴 Kogge-Stone 加法器的對於 N 位數目的實例完整加法器操作。在圖 6B 中, 第一 N 位運算元儲存於列 X 中且第二 N 位運算元儲存於列 Y 中。對於圖 6B 中描繪的實例加法操作, 計算中間項 $G_0、P_0、G_1、P_1、G_2、P_2、\cdots、G_{\log N+1}$ 以及 $P_{\log N+1}$ 。圖 6B 的最上區塊表示使用來自列 X 及 Y 的輸入運算元判定 G_0 及 P_0 的五個單獨操作。在第一操作中, 最上區塊判定列 X 的反相 (亦即, $\sim X$), 其儲存於列 1 中。第二操作判定列 Y 的反相 (亦即, $\sim Y$), 其儲存於列 2 中。第三操作判定操作列 X 反或列 Y , 其儲存於列 3 中。第四操作判定操作 $G_0 = \text{列 1 NOR 列 2}$, 其儲存於列 4 中。第五操作判定 $P_0 = \text{列 3 NOR 列 4}$, 其儲存於列 5 中。

【0035】 在圖 6B 的中間區塊中, 來自最上區塊的中間結果 G_0 及 P_0 用以判定中間結果 G_{i+1} 及 P_{i+1} , 其中 i 為行索引。亦即, 圖 6A 的最上區塊中判定的中間結果 G_0 及 P_0 用以判定中間結果 G_1 及 P_1 。中間結果 G_1 及 P_1 用以判定中間結果 G_2 及 P_2 等以判定中間結果 $G_{\log N+1}$ 及 $P_{\log N+1}$ 。在圖 6B 的最下區塊中, 結果列 $R1$ 及 $R2$ 分別儲存用於完整加法器操作的進位結果及求和結果。

【0036】 圖 6C 描繪可由 3T1C DRAM 計算胞元構形 201 提供的實例選擇器操作。列 1 儲存列 X 的反相 (亦即, $\sim X$) 的中間結果。列 2 儲存列 Y 的反相 (亦即, $\sim Y$) 的中間結果。列 3 儲存列 S 的反相 (亦即, $\sim S$) 的中間結果。列 4 儲存列 1 反或列 3 的中間結

果。列 5 儲存列 2 反或列 S 的中間結果。列 6 儲存列 4 反或列 5 的中間結果。列 R 儲存列 6 的反相的結果，亦即， $S \oplus X \oplus Y$ 。

【0037】圖 6D 描繪可由 3T1C DRAM 計算胞元構形 201 提供的替代實例選擇器操作。列 1 儲存列 X 的反相（亦即， $\sim X$ ）的中間結果。列 2 儲存列 S 的反相（亦即， $\sim S$ ）的中間結果。列 3 儲存列 1 反或列 3 的中間結果。列 4 儲存列 X 的反相（亦即， $\sim X$ ）的中間結果。列 R 儲存列 3 反或列 4 的結果，亦即， $S \oplus X \oplus \sim X$ 。

【0038】圖 6E 描繪可由 3T1C DRAM 計算胞元構形 201 提供的實例最大值/最小值（MIN）操作。列 1 儲存列 Y 的反相（亦即， $\sim Y$ ）的中間結果。列 2 儲存列 $X + (\sim Y + 1)$ 的中間結果。列 3 儲存 Cout > n 的中間結果。列 4 儲存 Cout > X:Y 的中間結果。列 R 儲存 MAX(X:Y) 的結果。

【0039】圖 6F 描繪可由 3T1C DRAM 計算胞元構形 201 提供的實例 1 位元乘法操作。列 1 儲存列 X 反或列 W 的中間結果。列 2 儲存列 X 反或列 1 的中間結果。列 3 儲存列 W 反或列 1 的中間結果。結果列 R 儲存列 2 反或列 3 的結果，亦即，列 X XNOR 列 W 的結果。

【0040】圖 6G 描繪可由 3T1C DRAM 計算胞元構形 201 提供的實例多位元乘法操作。在圖 6G 的上區塊中，列 1 儲存列 W 的反相（亦即， $\sim W$ ）的中間結果。列 2 儲存列 X 的反相經左移位 2^i 次（亦即， $\sim X \ll 2^i$ ）的中間結果，其中 i 為索引。列 3 儲存列 1 反或列 2 的中間結果，亦即， $PP_i = \sim W \text{ NOR } \sim X \ll 2^i$ 。在圖 6G 的下部區塊中，列 1 儲存列 PP_0 求和（SUM）列 PP_i （其為 $\sum PP_i$ ）的中間結果。列 2 儲存列 1 互斥反或列 W_{sign} 的中間結果。列 R 儲存 $X * W$ 的結

果。

【0041】 圖 7 描繪根據本文中所揭露主題的包含隨機資料區 715 的 DPU 700 的實例實施例的方塊圖。DPU 700 的具有與圖 1 中描繪的 DPU 100 的組件相同的參考指示符的各種組件是類似的且此處已省去此等類似組件的描述。DPU 700 的子陣列 102 包含隨機資料陣列 715 及轉換器至隨機陣列 716，以及（真實）資料胞元陣列 106、計算胞元陣列 107 以及記憶體陣列片內移位陣列 108。

【0042】 每一隨機資料陣列 715 可包含在至少一個行及至少一個列中配置的一或多個隨機計算胞元。在一個實施例中，隨機資料陣列 715 可包含 2K 行及 16 列。隨機資料陣列 715 中的行的數目與資料胞元陣列 106 及計算胞元陣列 107 中的行的數目相同。在另一實施例中，隨機資料陣列 715 可包含少於或多於 2K 行及/或少於或多於 16 列。在隨機資料陣列 715 中，使用「1」存在的機率且 2^n 位元用以表示 n 位元值。轉換器至隨機陣列 716 中的隨機數產生器可用於將實數轉換成隨機數。位 1 計數操作可用於將隨機數轉換回至實數。

【0043】 藉由使用隨機計算方法，加法可轉換成多工操作且乘法可轉換成及邏輯操作。舉例而言，圖 8A 描繪提供作為多工操作的隨機加法操作的電路，且圖 8B 描繪提供作為及邏輯操作的隨機乘法操作的電路。用於隨機計算的習知技術需要巨大記憶體容量；然而，本文中所揭露主題可用於提供高效隨機計算，此是因為基於 DRAM 的 DPU 能夠執行大並行 AND 及 MUX 操作。使用本文中所揭露的 DPU 的隨機計算亦使得有可能加速其中深度學習為典型應用的複雜操作。

【0044】 圖 9 描繪根據本文中所揭露主題的包含 DPU 的系統架構 900。系統架構 900 可包含硬體層 910、程式庫及驅動器層 920、框架層 930 以及應用層 940。

【0045】 硬體層 910 可包含具有嵌入的 DPU（諸如本文中所描述的 DPU）的硬體裝置及/或組件。裝置及/或組件的一個實施例可為可包含一或多個嵌入 DPU 的周邊組件高速互連（Peripheral Component Interconnect Express；PCIe）裝置 911。裝置及/或組件的另一實施例可為可包含一或多個嵌入 DPU 的雙直列記憶體模組（Dual In-line Memory Module；DIMM）912。應理解，系統架構 900 的硬體層 910 不限於 PCIe 裝置及/或 DIMM，但可包含系統單晶片（System on a Chip；SoC）裝置或可含有 DPU 的其他記憶體型裝置。可嵌入於硬體層 910 處的裝置及/或組件中的 DPU 可經組態為類似於圖 1 中的 DPU 100 及/或類似於圖 7 中的 DPU 700。在任一實施例中，DPU 的特定計算胞元陣列可經組態以包含 3T1C 計算胞元構形 201（圖 2A）或 1T1C 計算胞元構形 202（圖 2B）。

【0046】 系統架構 900 的程式庫及驅動器層 920 可包含 DPU 程式庫 921、DPU 驅動器 922 以及 DPU 編譯器 923。DPU 程式庫 921 可經組態以針對可在應用層 940 處操作的不同應用為用於硬體層 910 中的 DPU 中的每一子陣列提供最佳映射功能性、資源分配功能性以及排程功能性。

【0047】 在一個實施例中，DPU 程式庫 921 可為框架層 930 提供可包含諸如移動、加法、乘法等的操作的高層級應用程式設計介面（application programming interface；API）。舉例而言，DPU 程式庫 921 亦可包含用於標準型常式的實施，諸如但不限於可適用

於加速深度學習程序的前向卷積層及後向卷積層、池化層、正規化層，以及啟動層。在一個實施例中，DPU 程式庫 921 可包含類 API 功能，其映射用於卷積神經網路（convolution neural network；CNN）的整個卷積層的計算。另外，DPU 程式庫 921 可包含用於最佳化卷積層計算至 DPU 上的映射的類 API 功能。

【0048】 DPU 程式庫 921 亦可包含用於藉由將任務（批、輸出通道、像素、輸入通道、卷積核）內的任何個別或多個並行度映射成晶片、記憶體庫、子陣列及/或記憶體陣列片層級處的對應 DPU 並行度而最佳化資源分配的類 API 功能。另外，DPU 程式庫 921 可包含在初始化及/或運行時間處提供權衡效能（亦即，資料移動流動）及功率消耗的最佳 DPU 組態的類 API 功能。由 DPU 程式庫 921 提供的其他類 API 功能可包含設計旋鈕型功能，諸如設定每一記憶體庫的作用中子陣列的數目、每一作用中子陣列的輸入特徵圖的數目、特徵圖的分割，及/或卷積核的再次使用方案。其他類 API 功能可藉由分配子陣列的特定任務（諸如卷積計算、通道總計及/或資料分派）提供額外資源分配最佳化。若運算元將在整數與隨機數之間轉換，則 DPU 程式庫 921 包含在滿足精度約束條件的同時最小化額外負擔的類 API 功能。在精度低於預期情況下，DPU 程式庫 921 可包含再次使用用於隨機表示的額外位元計算值，或分流任務至其他硬體（諸如 CPU）的類 API 功能。

【0049】 DPU 程式庫 921 亦可包含同時排程 DPU 中的經啟動子陣列，並排程資料移動以使得其藉由計算操作隱藏的類 API 功能。

【0050】 DPU 程式庫 921 的另一態樣包含用於 DPU 進一步開發的延伸介面。在一個實施例中，DPU 程式庫 921 可提供用以使用反

或及移位邏輯直接程式化功能性以使得可提供除標準型操作（亦即，加法、乘法、MAX/MIN 等）以外的操作的介面。延伸介面亦可提供介面，以使得並非由 DPU 程式庫 921 特定支援的操作可在程式庫及驅動器層 920 處分擔至 SoC 控制器（未展示）、中央處理單元/圖形處理單元（central processing unit/graphics processing unit；CPU/GPU）組件及/或 CPU/張量處理單元（CPU/ Tensor Processing Unit；CPU/TPU）組件。DPU 程式庫 921 的又一態樣提供類 API 功能以在 DPU 記憶體不正用於計算時使用 DPU 的記憶體作為記憶體的延伸。

【0051】 DPU 驅動器 922 可經組態以提供在硬體層 910 處的 DPU、DPU 程式庫 921 以及在較高層處的作業系統（operating system；OS）之間的介面連接以將 DPU 硬體層整合至系統中。亦即，DPU 驅動器 922 將 DPU 曝露於系統 OS 及 DPU 程式庫 921。在一個實施例中，DPU 驅動器 922 可在初始化時提供 DPU 控制。在一個實施例中，DPU 驅動器 922 可發送呈 DRAM 型位址或 DRAM 型位址的序列形式的指令至 DPU 且可控制至 DPU 中及自 DPU 的資料移動。除了處置 DPU 至 CPU 及/或 DPU 至 GPU 通信之外，DPU 驅動器 922 可提供多 DPU 通信。

【0052】 DPU 編譯器 923 可將來自 DPU 程式庫 921 的 DPU 碼編譯成呈由 DPU 驅動器 922 使用以控制 DPU 的記憶體位址形式的 DPU 指令。由 DPU 編譯器 923 產生的 DPU 指令可為在 DPU 中的一個及/或兩個列上操作的單個指令；向量指令，及/或收集的向量、操作讀取指令。

【0053】 框架層 930 可經組態以向程式庫及驅動器層 920 以及硬

體層 910 提供使用者友好介面。在一個實施例中，框架層 930 可提供與在應用層 940 處的廣泛範圍的應用相容且使 DPU 硬體層 910 對於使用者透明的使用者友好介面。在另一實施例中，框架層 930 可包括添加量化功能至現有習知方法的框架延伸部，諸如但不限於 Torch7 型應用及 TensorFlow 型應用。在一個實施例中，框架層 930 可包括添加量化功能至訓練演算法。在另一實施例中，框架層 930 可提供對現有除法、乘法以及平方根的批正規化方法的覆寫，以成為除法、乘法以及平方根的移位近似方法。在再一實施例中，框架層 930 可提供允許使用者設定用於計算的位元數目的延伸。在另一實施例中，框架層 930 提供包覆來自 DPU 程式庫及驅動器層 920 的多 DPU API 至框架層 930 的能力，以使得使用者可類似於多個 GPU 的使用而使用硬體層處的多個 DPU。框架層 930 的再一特徵允許使用者指派功能給在硬體層 910 處的 DPU 或 GPU。

【0054】 應用層 940 可包含廣泛範圍的應用，諸如但不限於影像標籤處理、自駕駛/引導車輛、AlphaGo 型深度頭腦應用及/或話音研究。

【0055】 如將由熟習此項技術者認識到，可跨廣泛範圍的應用修改本文中所描述的新穎概念且使其變化。因此，所主張主題的範疇不應限於上文所論述的特定例示性教示中的任一者，而實際上由以下申請專利範圍界定。

【符號說明】

【0056】

- 100、700：DPU
- 101a、101b：記憶體庫
- 102a、102b：子陣列
- 103：緩衝器
- 104：系統匯流排
- 105、105a、105b、105n：記憶體陣列片
- 106：資料胞元陣列
- 107：計算胞元陣列
- 107a、107b、107c、107d：計算胞元
- 108：記憶體陣列片內移位陣列
- 109：虛線
- 110：資料胞元陣列解碼器
- 111：計算胞元陣列解碼器
- 112e、112f：資料移位線
- 112：記憶體陣列片間移位陣列
- 113：記憶體陣列片間轉遞陣列
- 113g：第一資料轉遞線
- 113h：第二資料轉遞線
- 114：子陣列控制器
- 201：三電晶體一電容器（3T1C） DRAM 計算胞元構形
- 202：一電晶體一電容器（1T1C） DRAM 計算胞元構形
- 715：隨機資料陣列
- 716：轉換器至隨機陣列
- 900：系統架構

910：硬體層

911：周邊組件高速互連（PCIe）裝置

912：雙直列記憶體模組（DIMM）

920：程式庫及驅動器層

921：DPU 程式庫

922：DPU 驅動器

923：DPU 編譯器

930：框架層

940：應用層

addr：位址

ALU：算術邏輯單元

BL：位元線

C₁、C₂：電容器

CNTL1、CNTL2、CNTL3、CNTL4：控制信號

FCL：轉遞控制線

FDL：資料轉遞線

FSL：轉遞區段線

ISLcL：記憶體陣列片間移位控制線

LATCH：栓鎖器

MUX：多工器

SA：感測放大器

SL：移位線

SLcL：左移位控制線

SML：移位遮罩線

SRcL：右移位控制線

T₁：第一電晶體

T₂：第二電晶體

T₃：第三電晶體

T₄、T₅、T₆、T_{112a}、T_{112b}、T_{112c}、T_{112d}、T_{113a}、T_{113b}、T_{113c}、

T_{113d}、T_{113e}、T_{113f}：電晶體

WEN_X、WEN_Y、WEN_R：寫入啟用線

WL、WL_R、WL_X、WL_Y：字線

【發明申請專利範圍】

【第1項】 一種DPU操作作用的系統，包括：

介面，用以接收命令；

程式庫，用以判定對應於所接收的所述命令的至少一個動態隨機存取記憶體（DRAM）處理單元（DPU）操作；

編譯器，用以形成對應於所接收的所述命令的經判定所述至少一個 DPU 操作的至少一個 DPU 指令；以及

驅動器，其用以發送所述至少一個 DPU 指令至至少一個 DPU，所述 DPU 包括至少一個計算胞元陣列，所述至少一個計算胞元陣列包括以具有至少一個行的陣列方式配置的多個基於 DRAM 的計算胞元，所述至少一個行包括至少三個列的基於 DRAM 的計算胞元，所述至少三個列的基於 DRAM 的計算胞元經組態以提供在所述至少三個列的第一列及第二列上操作的邏輯功能且經組態以在所述至少三個列的第三列中儲存所述邏輯功能的結果。

【第2項】 如申請專利範圍第1項所述的系統，其中所述至少一個行的所述基於DRAM的計算胞元各自包括三電晶體及一電容器（3T1C）式DRAM記憶體胞元，或一電晶體及一電容器（1T1C）式DRAM記憶體胞元，且

其中所述至少一個行的所述基於 DRAM 的計算胞元提供反或邏輯功能。

【第3項】 如申請專利範圍第2項所述的系統，其中所述DPU更包括以具有至少一個行的陣列方式配置的多個基於DRAM的隨機計算胞元，所述至少一個行包括至少三個列的基於DRAM的隨機計算胞元，所述至少三個列的基於DRAM的隨機計算胞元經組態以

提供在所述至少三個列的第一列及第二列上操作的隨機邏輯功能且經組態以在所述至少三個列的第三列中儲存所述隨機邏輯功能的結果。

【第4項】如申請專利範圍第3項所述的系統，其中所述至少一個行的所述基於DRAM的隨機計算胞元各自包括三電晶體及一電容器（3T1C）式DRAM記憶體胞元，或一電晶體及一電容器（1T1C）式DRAM記憶體胞元。

【第5項】如申請專利範圍第4項所述的系統，其中所述至少一個DPU操作包括隨機計算操作。

【第6項】如申請專利範圍第1項所述的系統，其中所述DPU更包括以具有至少一個行的陣列方式配置的多個基於DRAM的隨機計算胞元，所述至少一個行包括至少三個列的基於DRAM的隨機計算胞元，所述至少三個列的基於DRAM的隨機計算胞元經組態以提供在所述至少三個列的第一列及第二列上操作的隨機邏輯功能且經組態以在所述至少三個列的第三列中儲存所述隨機邏輯功能的結果。

【第7項】如申請專利範圍第1項所述的系統，其中所述程式庫進一步回應於所接收的所述命令而判定整個卷積層的映射。

【第8項】如申請專利範圍第1項所述的系統，其中所述程式庫進一步判定所述至少一個DPU操作內的多個並行度，所述多個並行度映射多個DPU以執行對應於所接收的所述命令的所述至少一個DPU操作。

【第9項】如申請專利範圍第1項所述的系統，其中所述驅動器進一步基於所接收的所述命令控制至所述DPU中的資料移動及自所

述DPU的資料移動。

【第10項】 如申請專利範圍第1項所述的系統，其中所述編譯器進一步形成對應於所接收的所述命令的在DPU的至少一個列上操作的單個DPU指令。

【第11項】 一種DPU操作作用的系統，包括：

至少一個動態隨機存取記憶體（DRAM）處理單元（DPU），
每一 DPU 包括：

至少一個計算胞元陣列，包括以具有至少一個行的陣列方式配置的多個基於 DRAM 的計算胞元，所述至少一個行包括至少三個列的基於 DRAM 的計算胞元，所述至少三個列的基於 DRAM 的計算胞元經組態以提供在所述至少三個列的第一列及第二列上操作的真實邏輯功能且經組態以在所述至少三個列的第三列中儲存所述真實邏輯功能的結果；以及

多個基於 DRAM 的隨機計算胞元，以具有至少一個行的陣列方式配置，所述至少一個行包括至少三個列的基於 DRAM 的隨機計算胞元，所述至少三個列的基於 DRAM 的隨機計算胞元經組態以提供在所述至少三個列的第一列所接收的第一串流資料及所述至少三個列的第二列所接收的第二串流資料上操作的隨機邏輯功能且經組態以在所述至少三個列的第三列中儲存所述隨機邏輯功能所得到的串流資料；

介面，用以接收命令；

程式庫，用以判定對應於所接收的所述命令的至少一個 DPU 指令；以及

驅動器，用以發送對應於所接收的所述命令的所述至少一個

DPU 指令至所述至少一個 DPU。

【第12項】 如申請專利範圍第11項所述的系統，其中所述至少一個行的所述基於DRAM的計算胞元各自包括三電晶體及一電容器（3T1C）式DRAM記憶體胞元，或一電晶體及一電容器（1T1C）式DRAM記憶體胞元，且

其中所述至少一個行的所述基於 DRAM 的隨機計算胞元各自包括三電晶體及一電容器（3T1C）式 DRAM 記憶體胞元，或一電晶體及一電容器（1T1C）式 DRAM 記憶體胞元。

【第13項】 如申請專利範圍第12項所述的系統，其中所述至少一個行的所述基於DRAM的計算胞元提供反或邏輯功能。

【第14項】 如申請專利範圍第11項所述的系統，其中所述至少一個DPU操作包括隨機計算操作。

【第15項】 如申請專利範圍第11項所述的系統，更包括編譯器，所述編譯器進一步形成對應於所接收的所述命令的在DPU的至少一個列上操作的單個DPU指令。

【第16項】 一種DPU操作作用的系統，包括：

至少一個動態隨機存取記憶體（DRAM）處理單元（DPU），包括至少一個隨機計算胞元陣列，所述至少一個隨機計算胞元陣列包括以具有至少一個行的陣列方式配置的多個基於 DRAM 的隨機計算胞元，所述至少一個行包括至少三個列的基於 DRAM 的隨機計算胞元，所述至少三個列的基於 DRAM 的隨機計算胞元經組態以提供在所述至少三個列的第一列所接收的第一串流資料及所述至少三個列的第二列所接收的第二串流資料上操作的隨機邏輯功能且經組態以在所述至少三個列的第三列中儲存所述隨機邏輯

功能所得到的串流資料；

程式庫，用以判定對應於接收命令的至少一個隨機 DPU 指令；
以及

驅動器，用以發送對應於所述接收命令的所述至少一個隨機 DPU 指令至包括所述多個基於 DRAM 的隨機計算胞元的所述至少一個 DPU。

【第17項】 如申請專利範圍第16項所述的系統，其中所述至少一個行的所述基於DRAM的隨機計算胞元各自包括三電晶體及一電容器(3T1C)式DRAM記憶體胞元，或一電晶體及一電容器(1T1C)式DRAM記憶體胞元。

【第18項】 如申請專利範圍第16項所述的系統，其中所述驅動器進一步基於所述接收命令控制至包括所述多個基於DRAM的隨機計算胞元的所述DPU中的資料移動及自所述DPU的資料移動。

【第19項】 如申請專利範圍第16項所述的系統，更包括至少一個計算胞元陣列，所述至少一個計算胞元陣列包括以具有至少一個行的陣列方式配置的多個基於DRAM的計算胞元，所述至少一個行包括至少三個列的基於DRAM的計算胞元，所述至少三個列的基於DRAM的計算胞元經組態以提供在所述至少三個列的第一列及第二列上操作的真實邏輯功能且經組態以在所述至少三個列的第三列中儲存所述真實邏輯功能的結果，

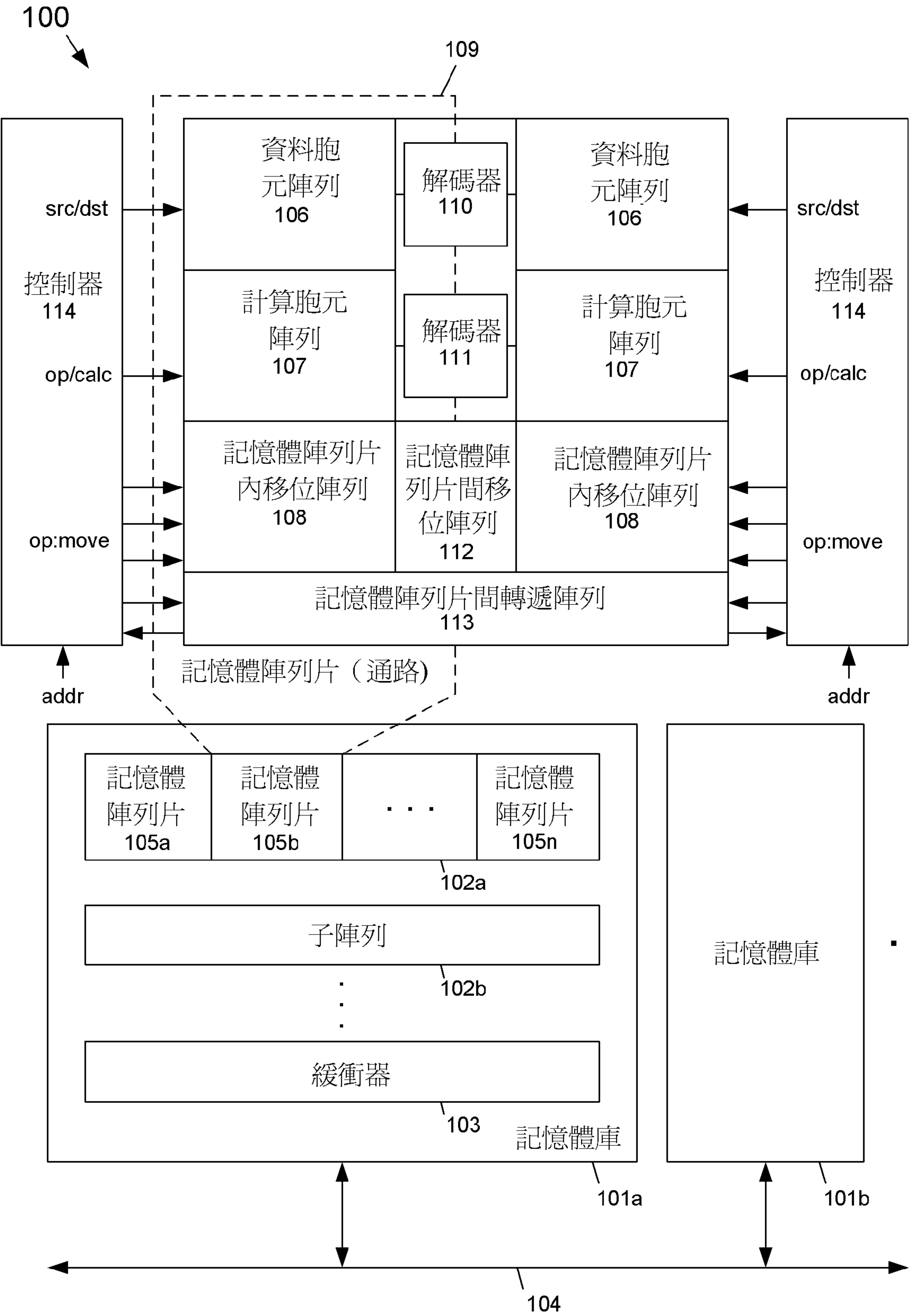
其中所述程式庫進一步判定用於所述至少一個計算胞元陣列的對應於第二接收命令的至少一個 DPU 指令；以及

驅動器，用以發送對應於所述第二接收命令的所述至少一個 DPU 指令至包括所述多個基於 DRAM 的計算胞元的所述至少一個

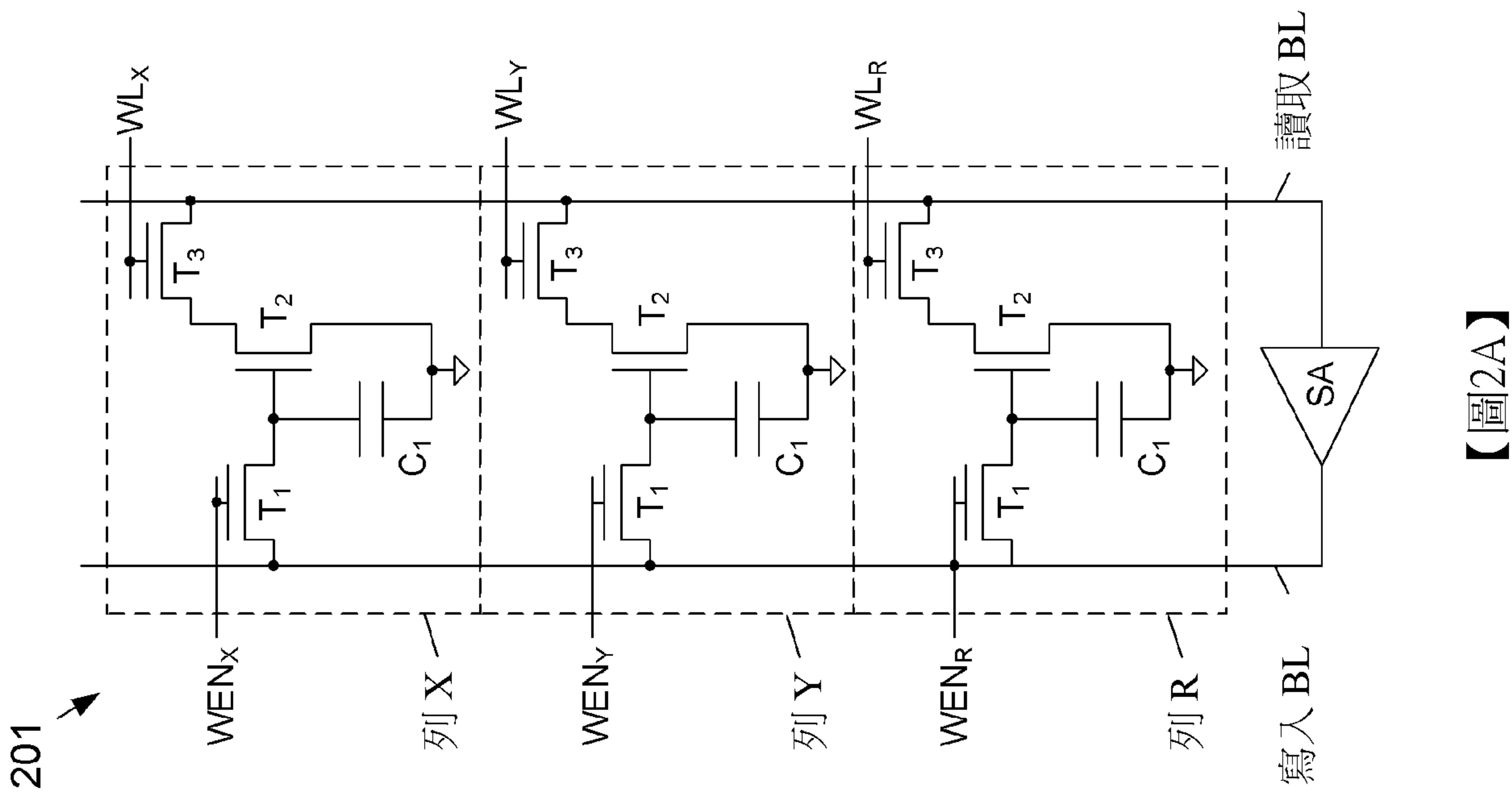
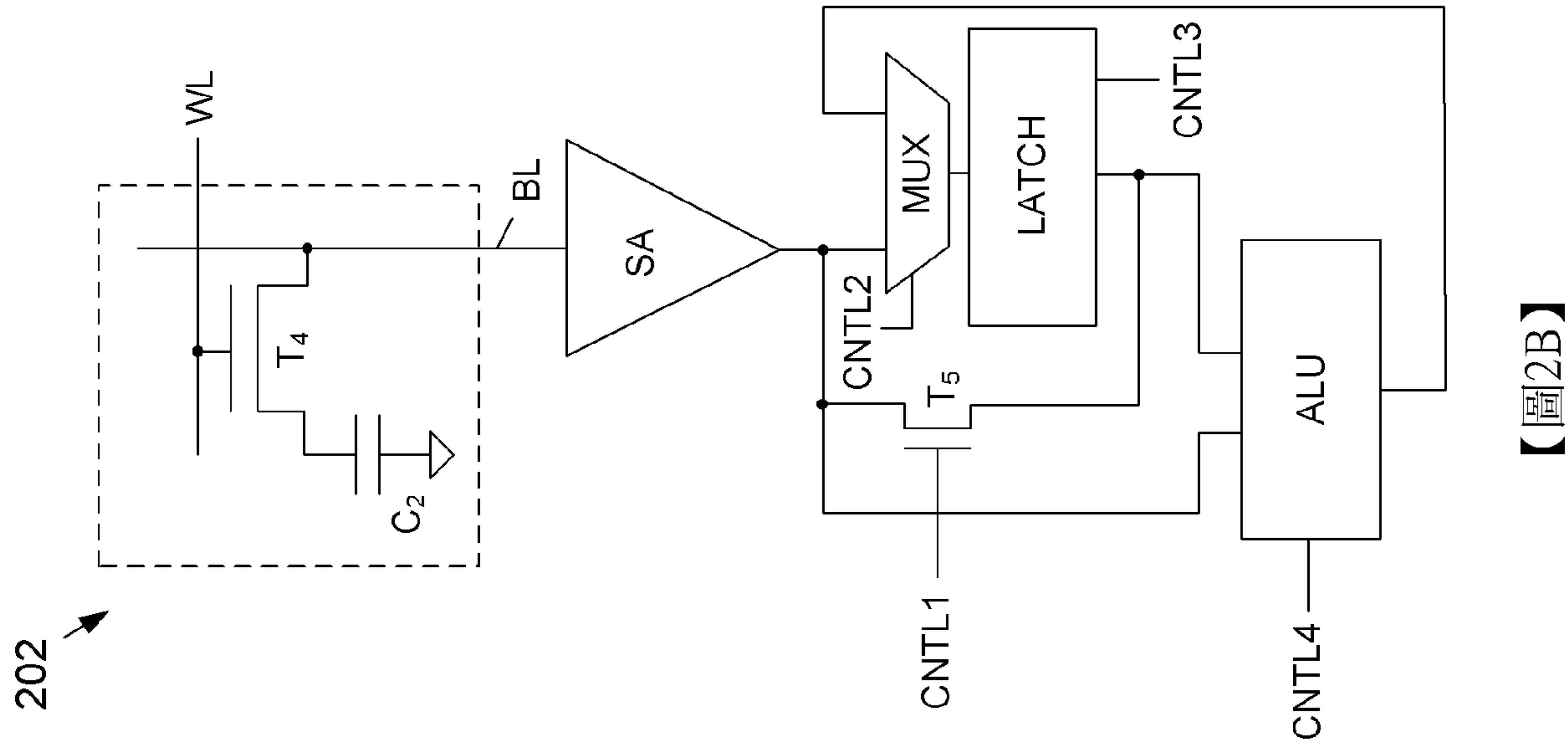
計算胞元陣列。

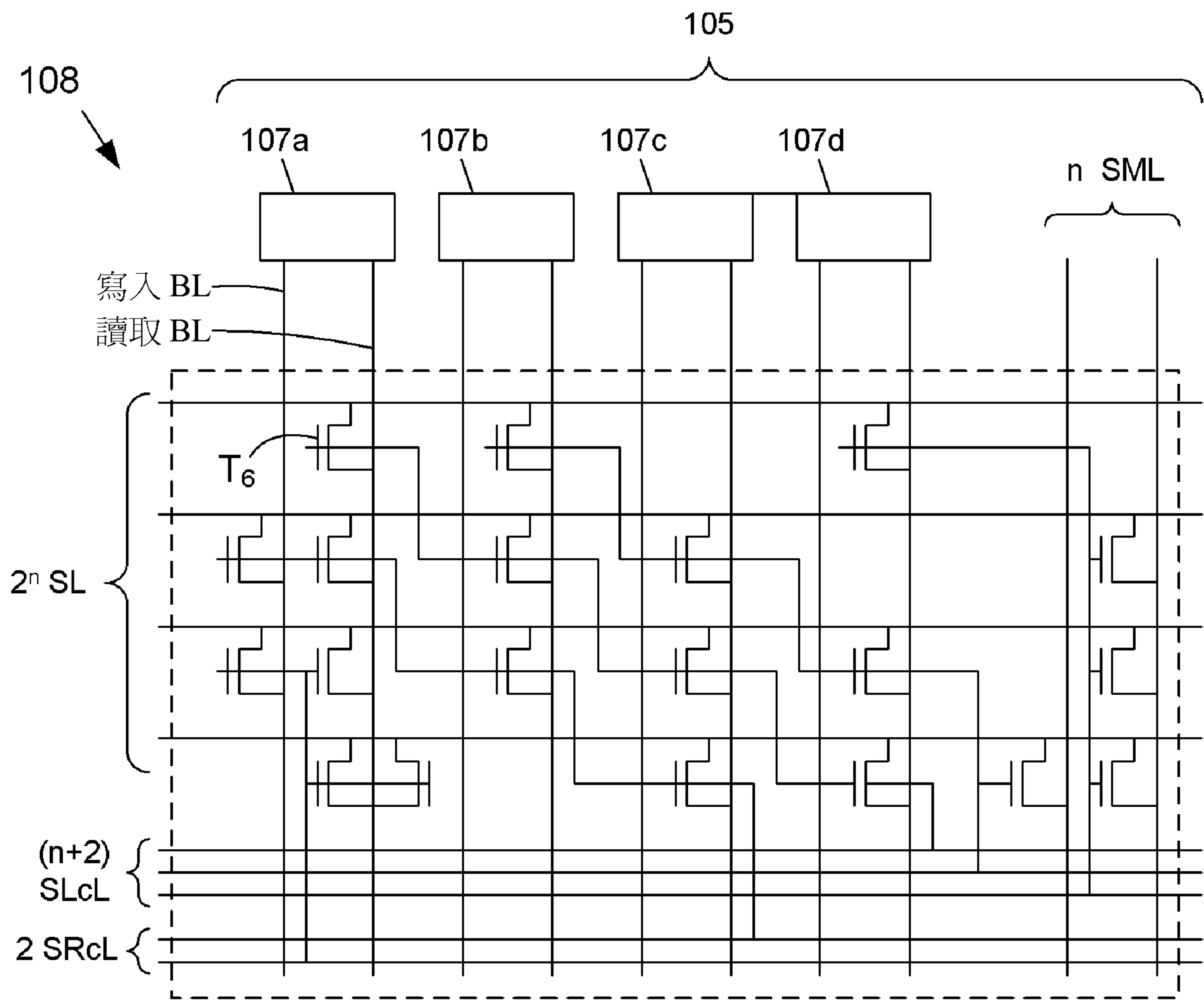
【第20項】 如申請專利範圍第19項所述的系統，其中所述至少一個行的所述基於DRAM的隨機計算胞元各自包括三電晶體及一電容器(3T1C)式DRAM記憶體胞元，或一電晶體及一電容器(1T1C)式DRAM記憶體胞元。

【發明圖式】

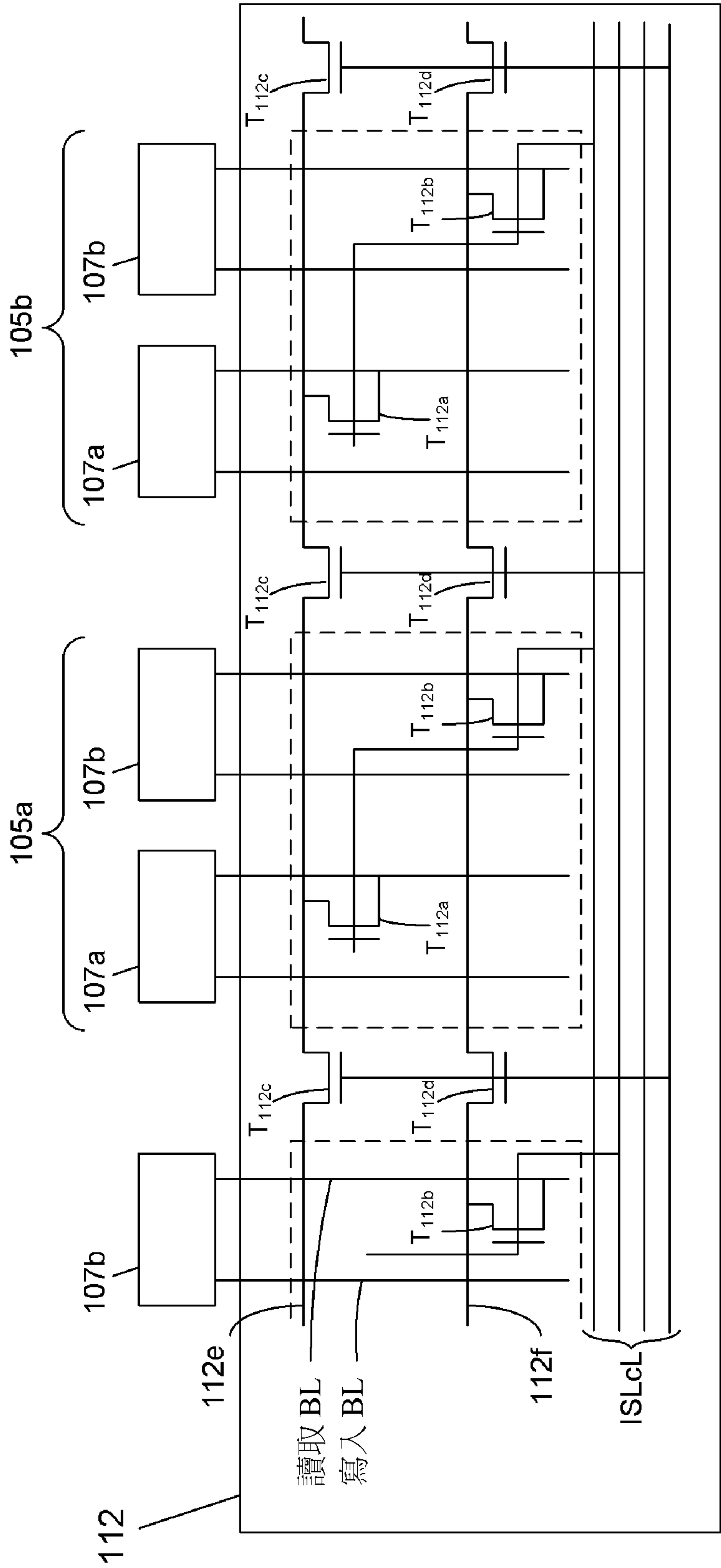


【圖1】

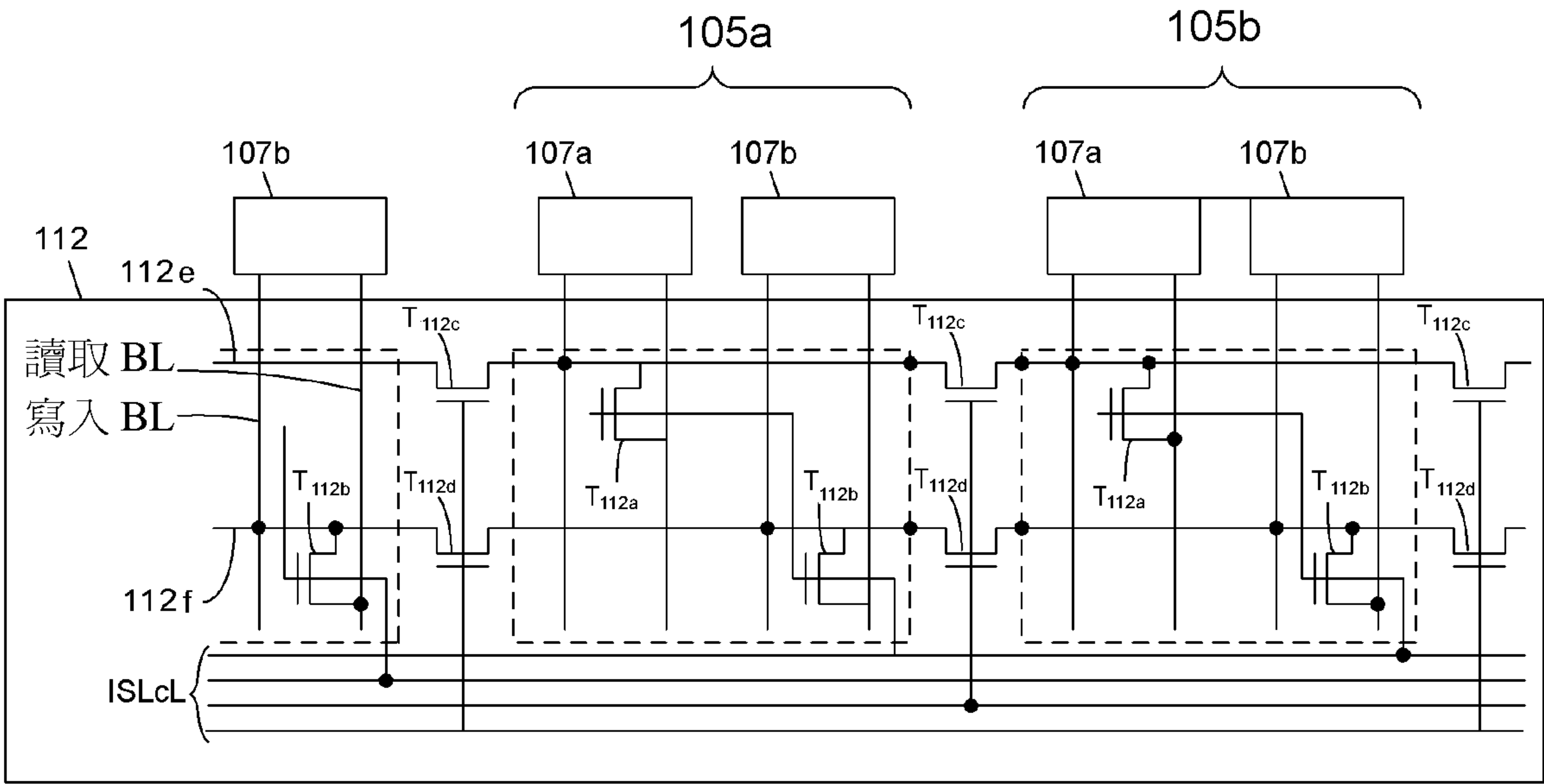




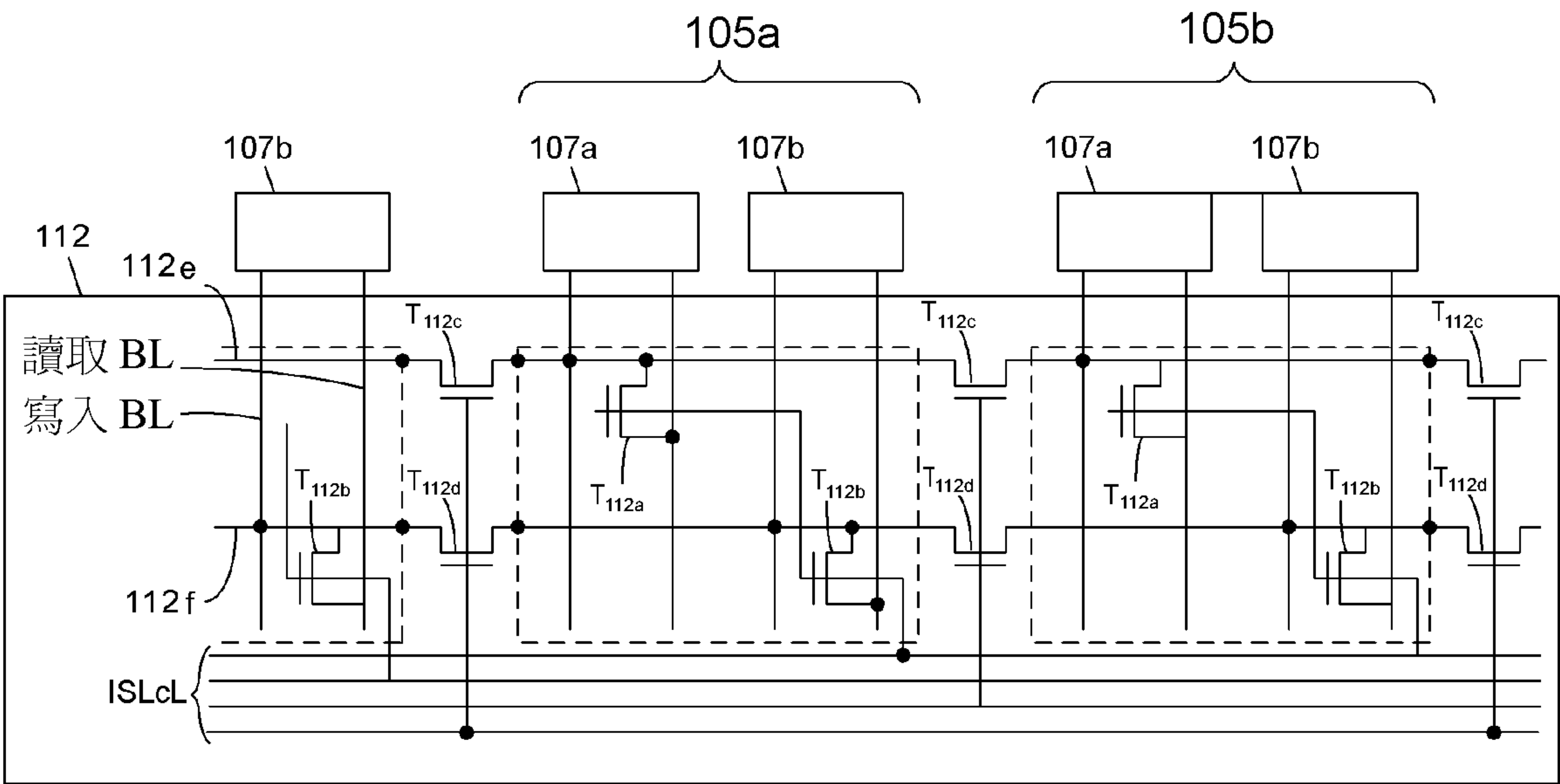
【圖3】



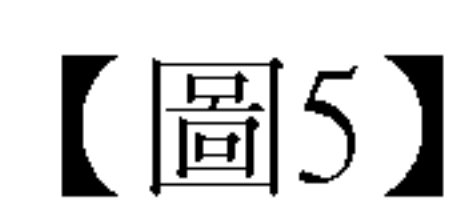
【圖4A】

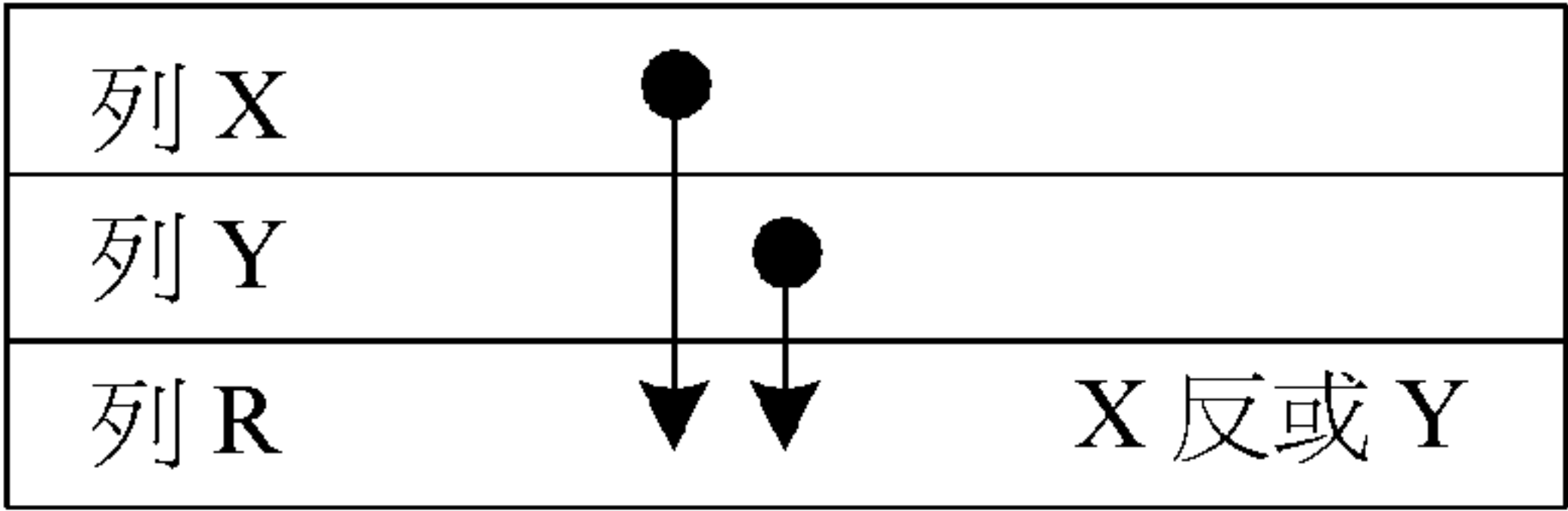


【圖4B】

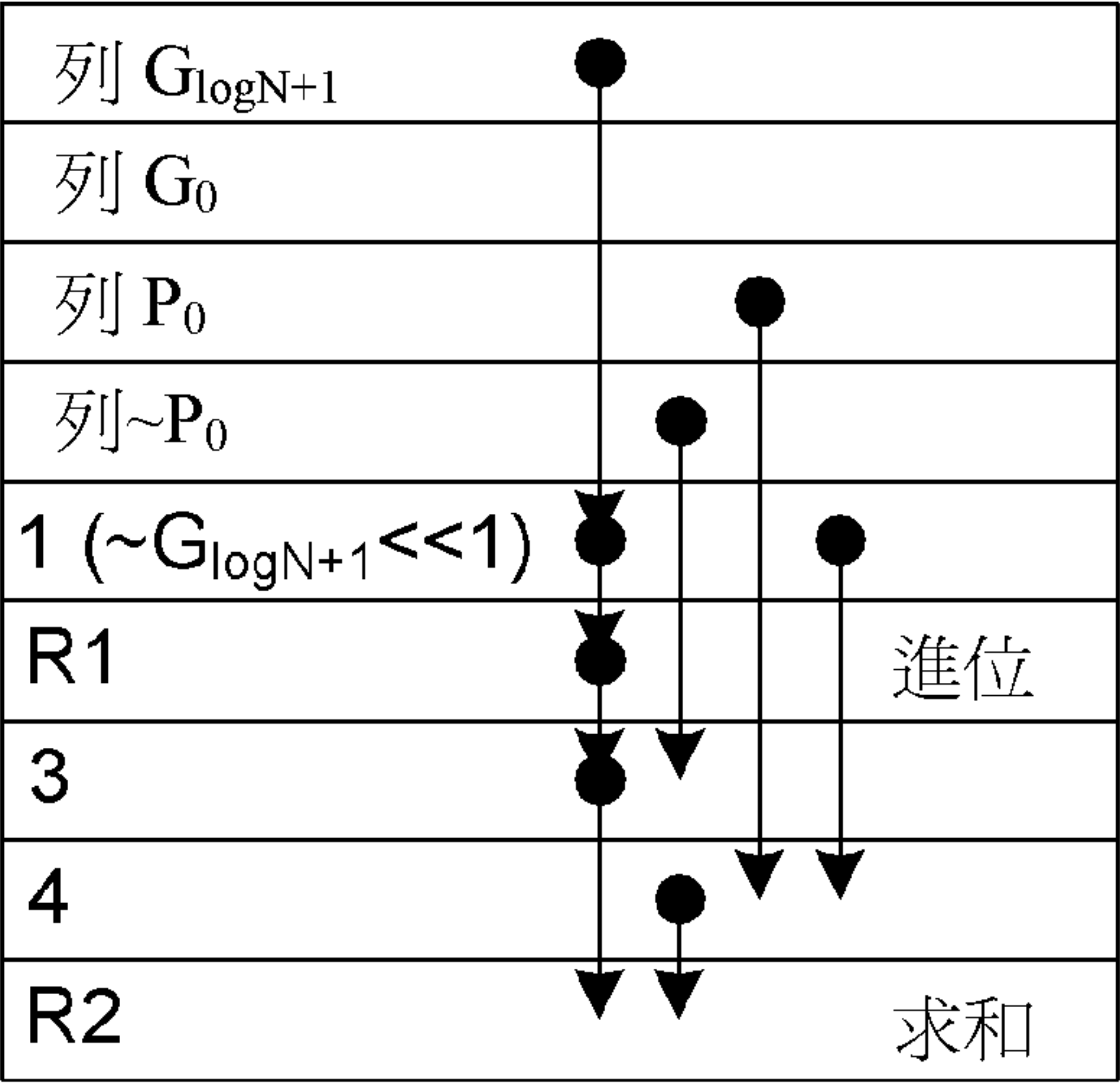
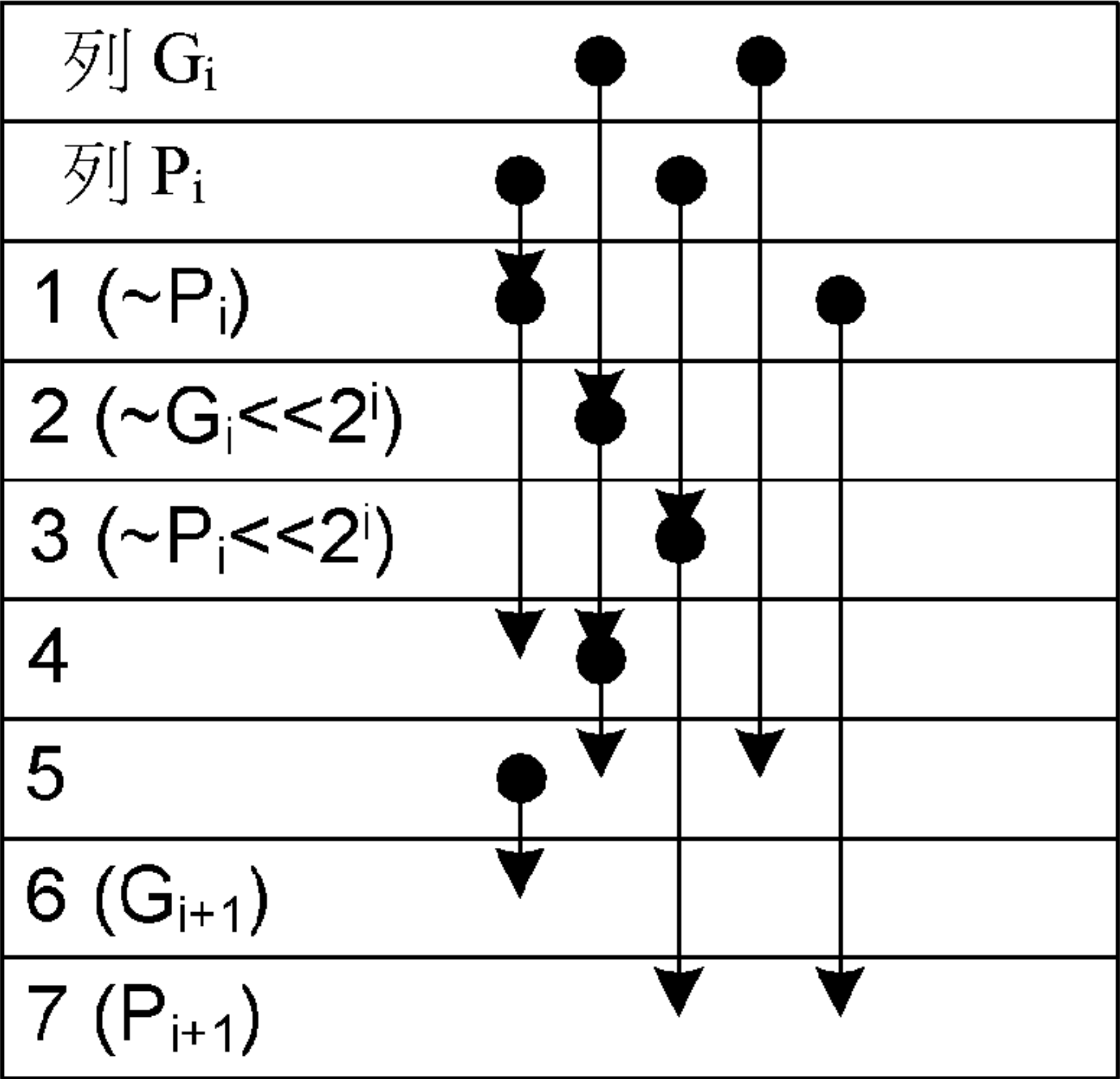
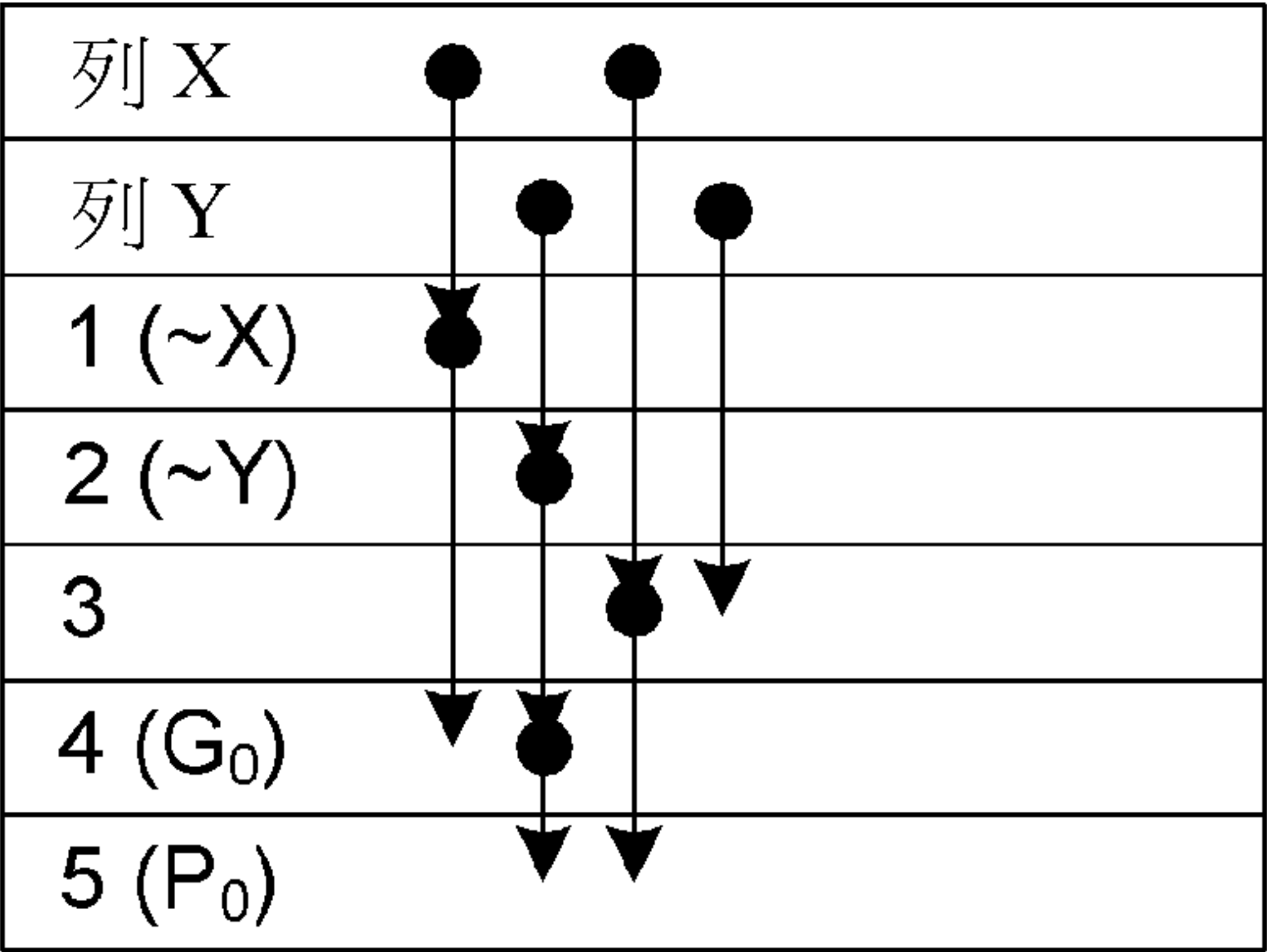


【圖4C】

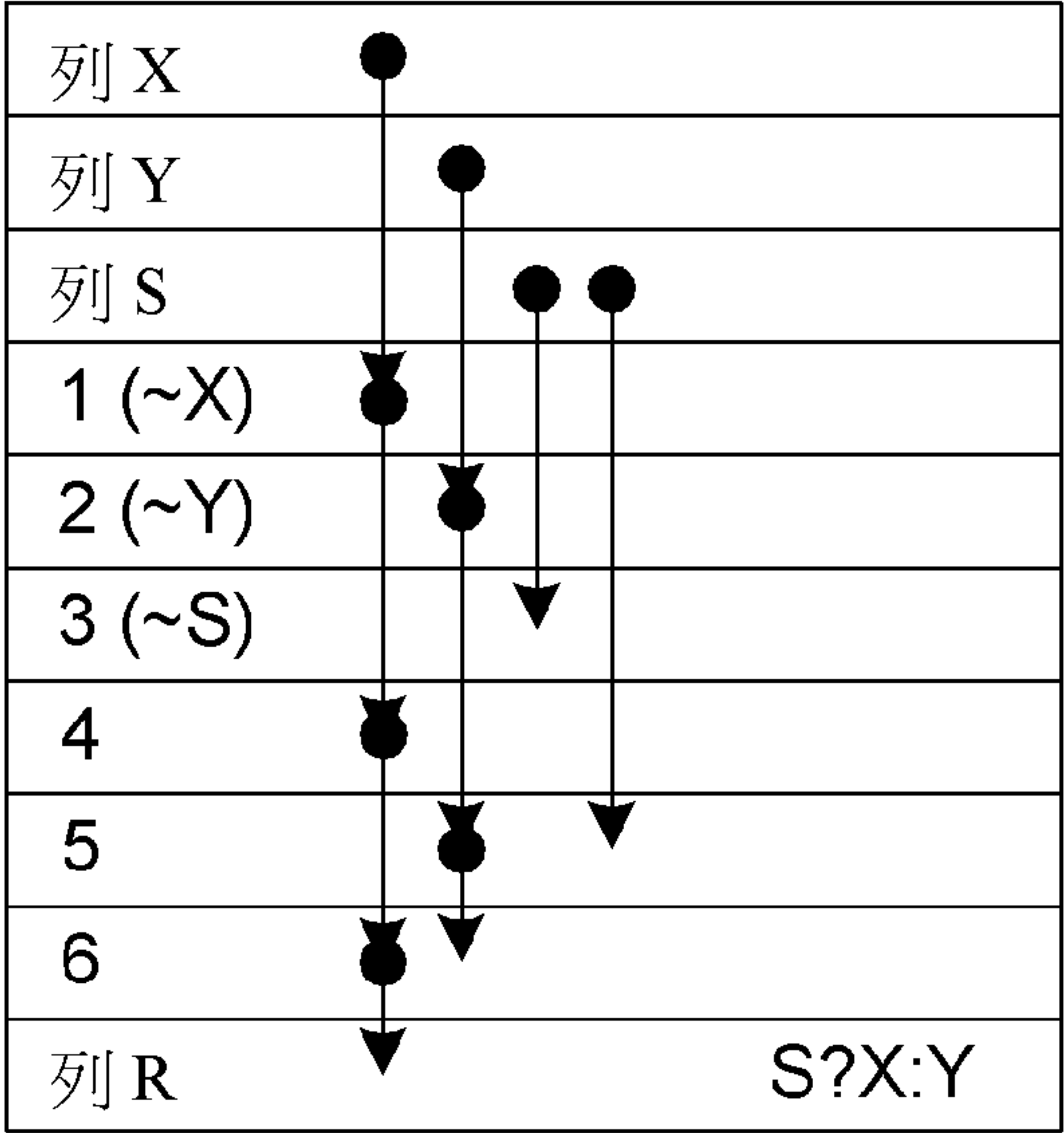




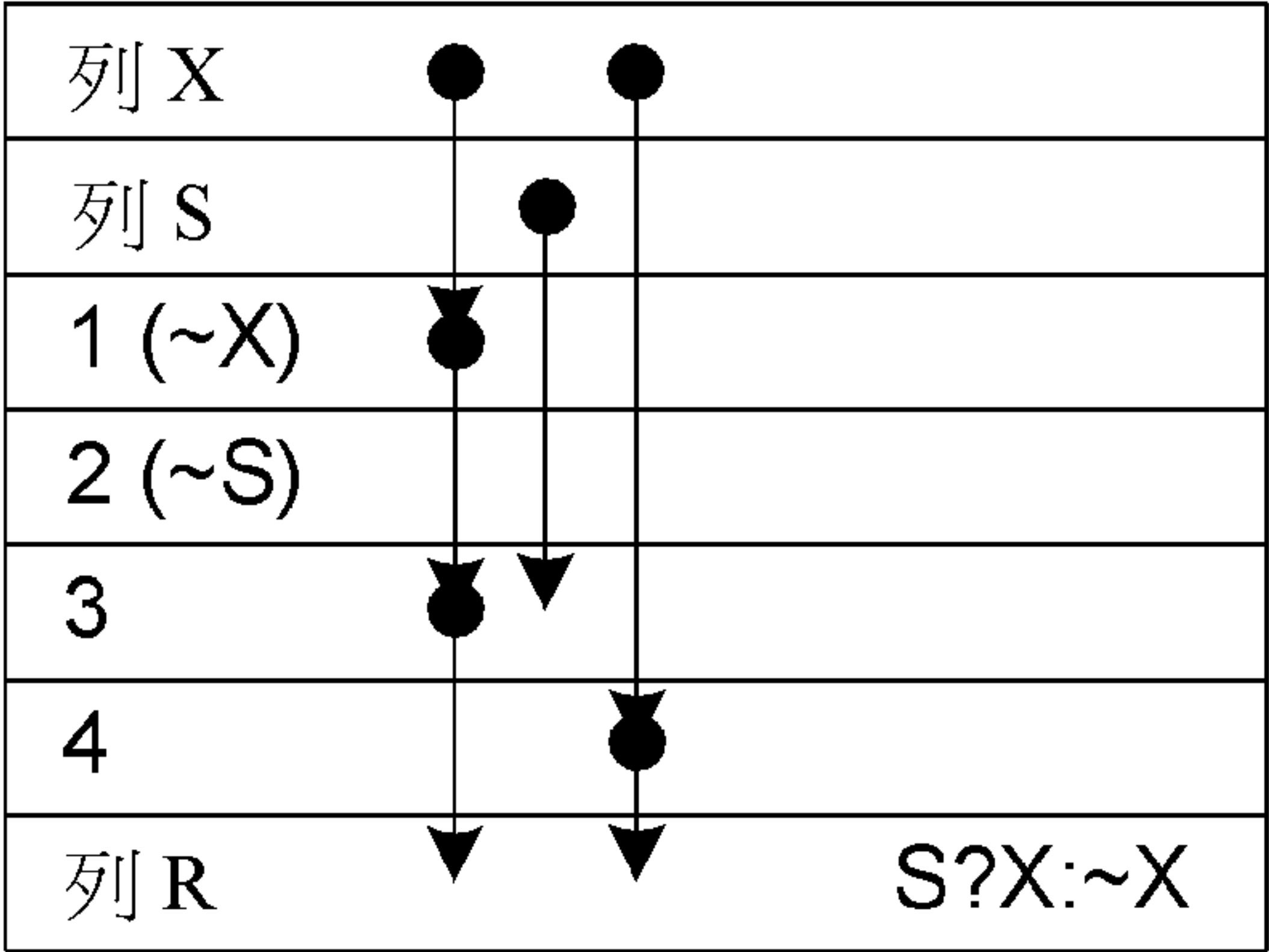
【圖6A】



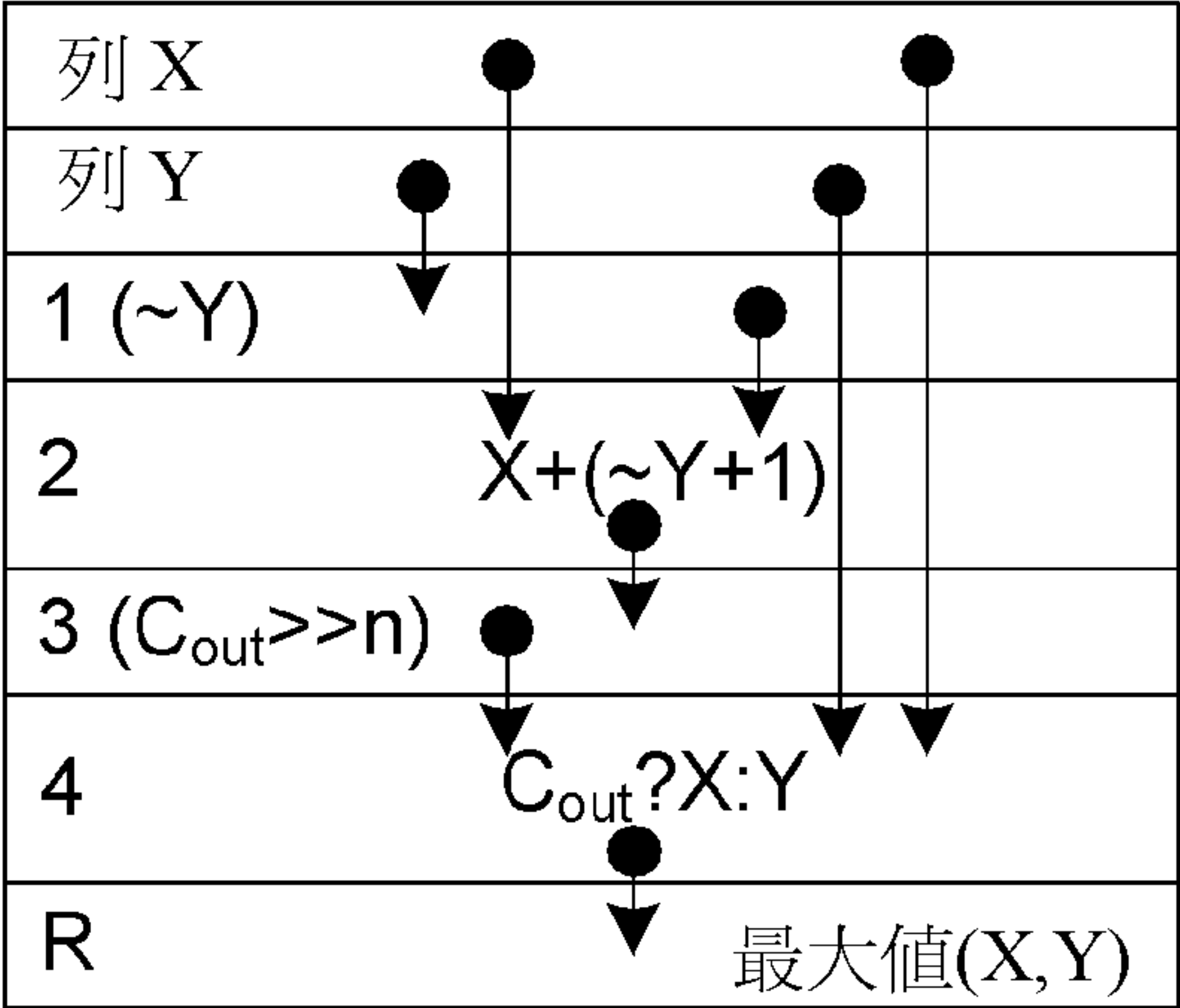
【圖6B】



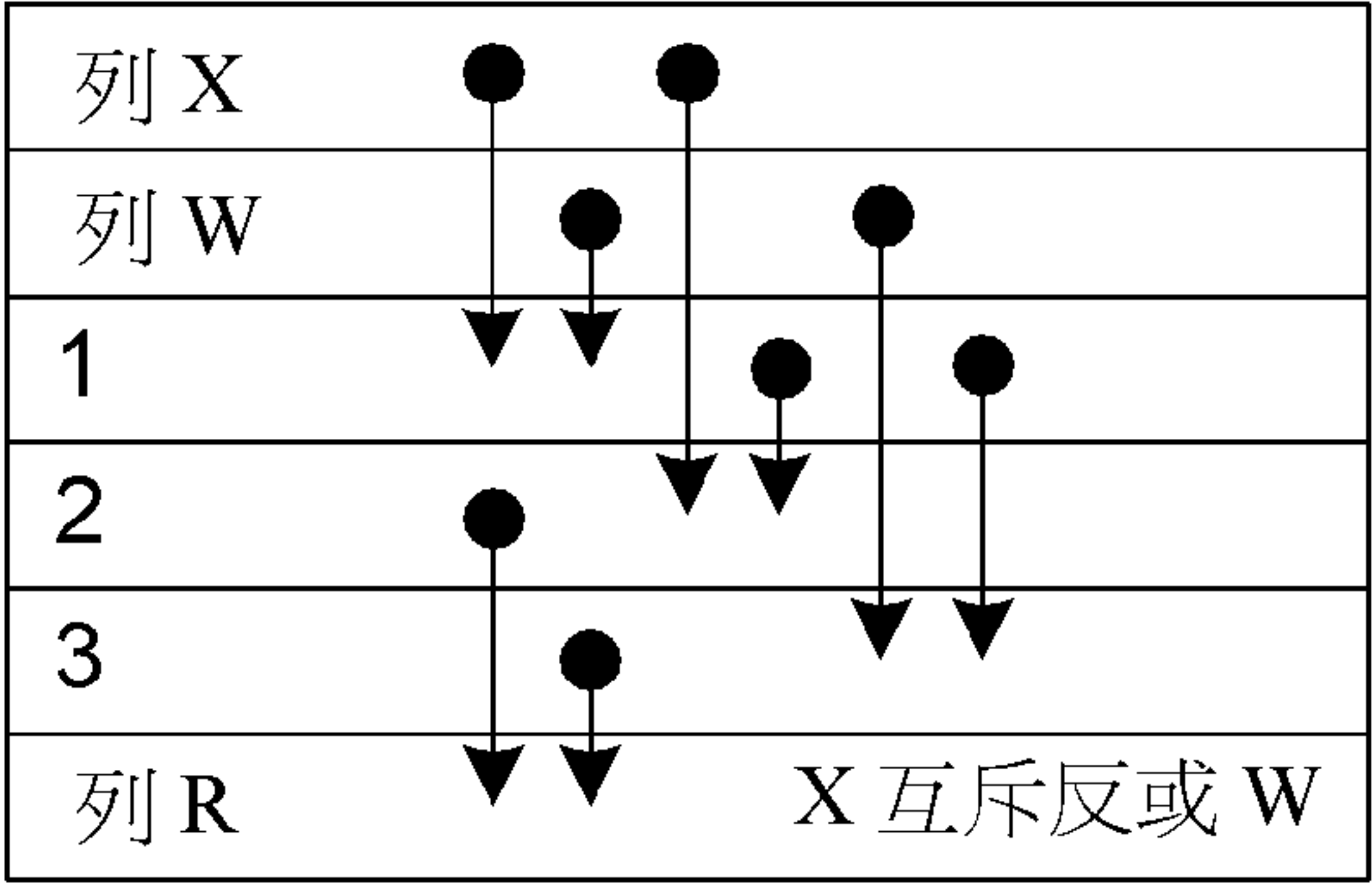
【圖6C】



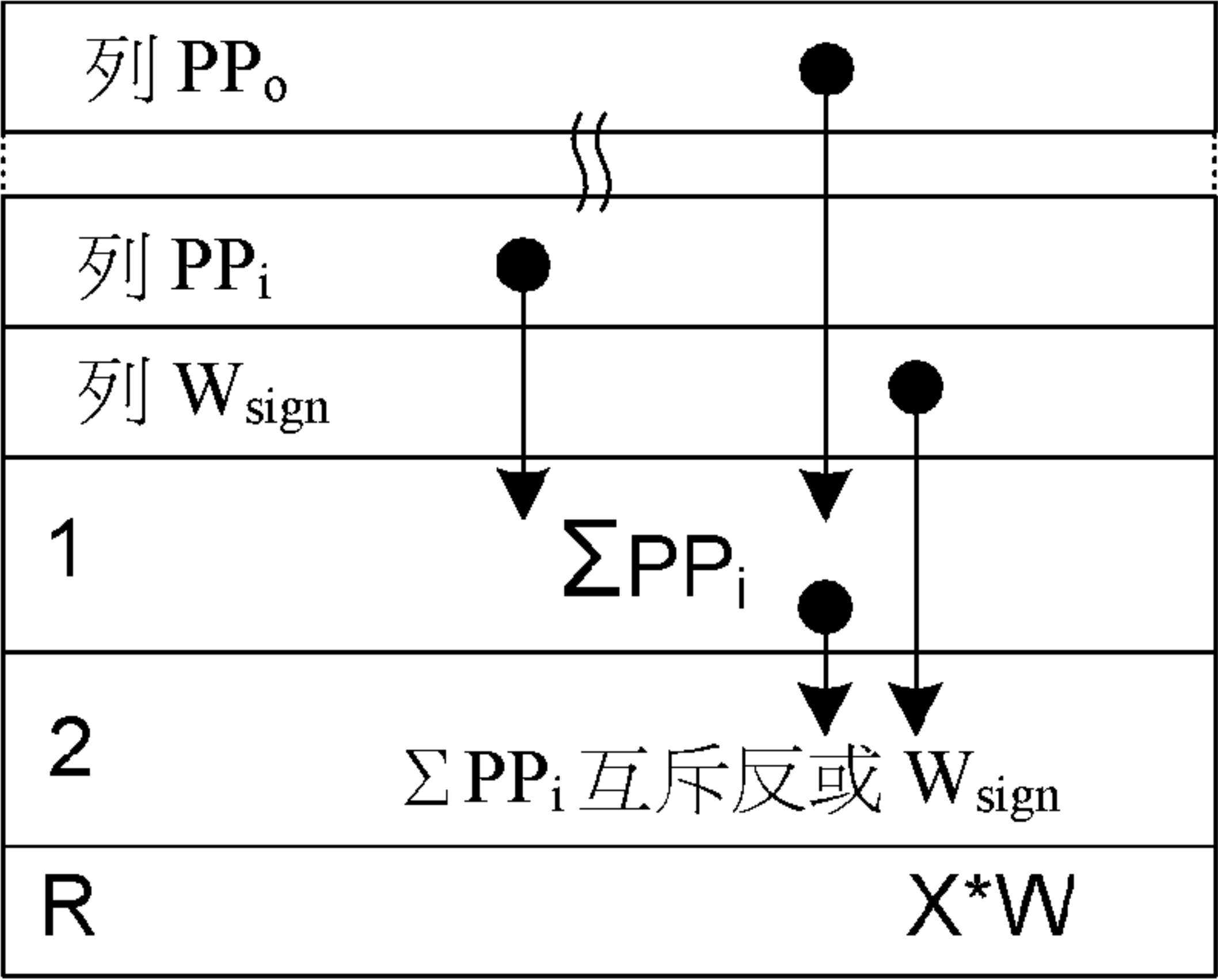
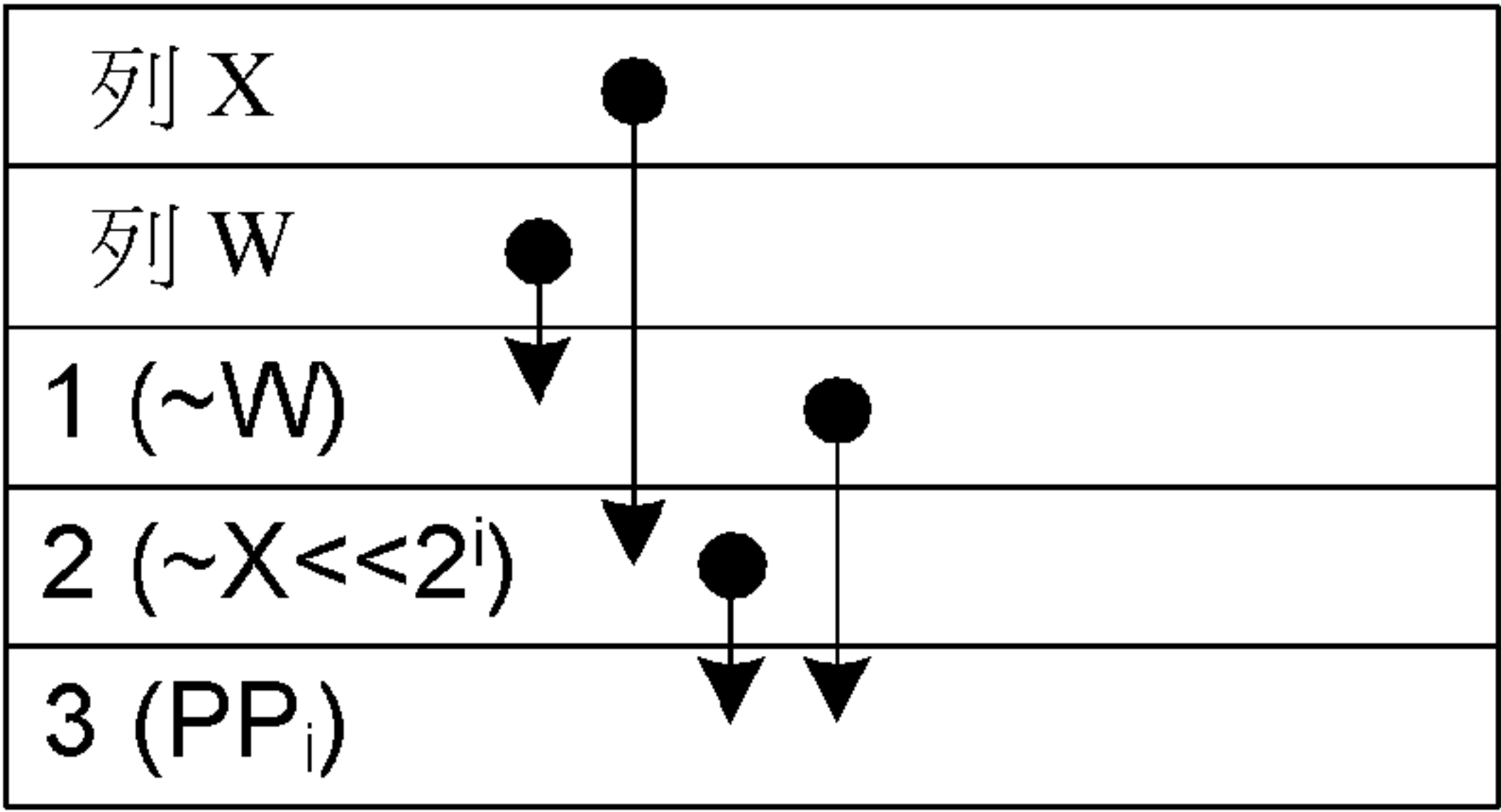
【圖6D】



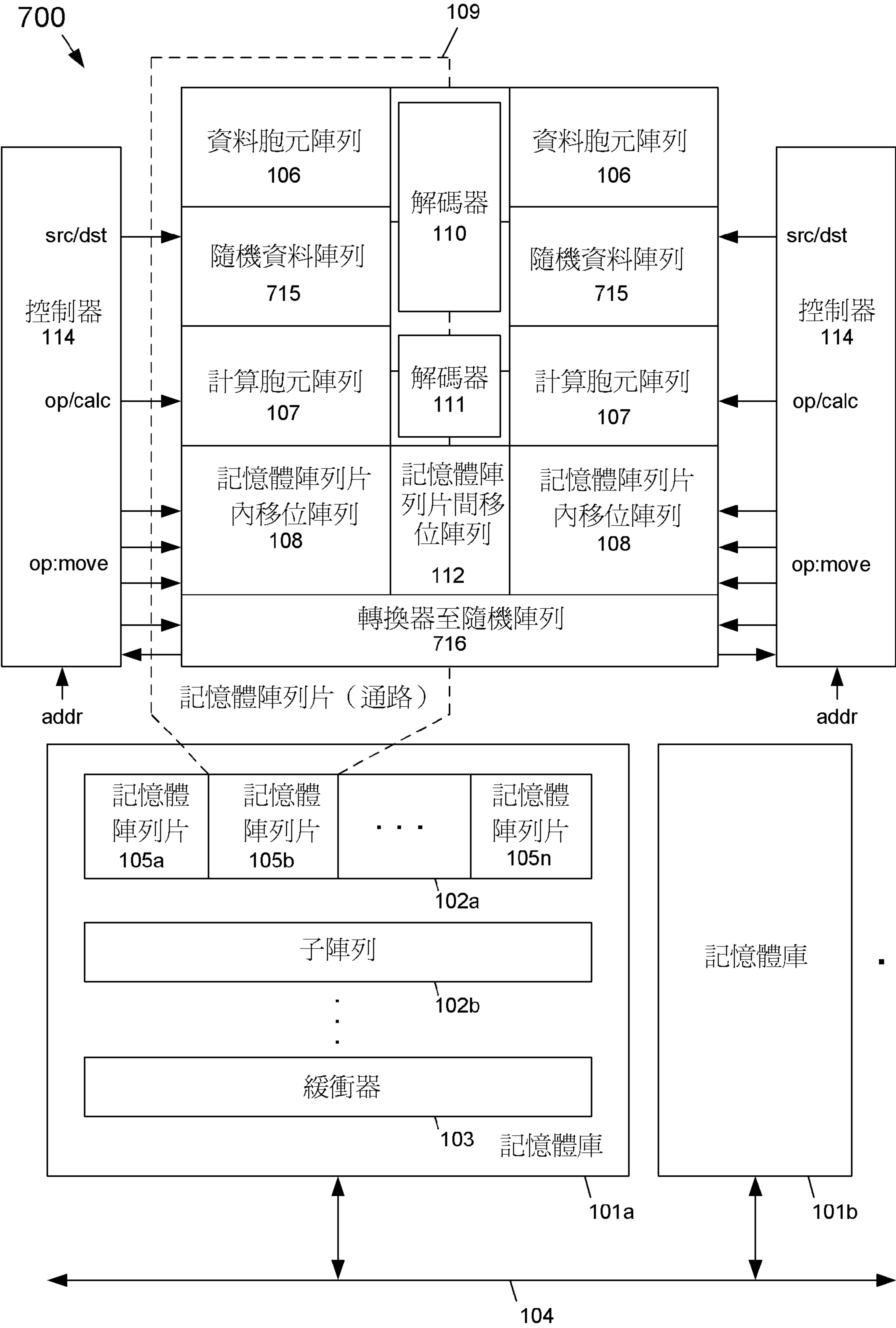
【圖6E】



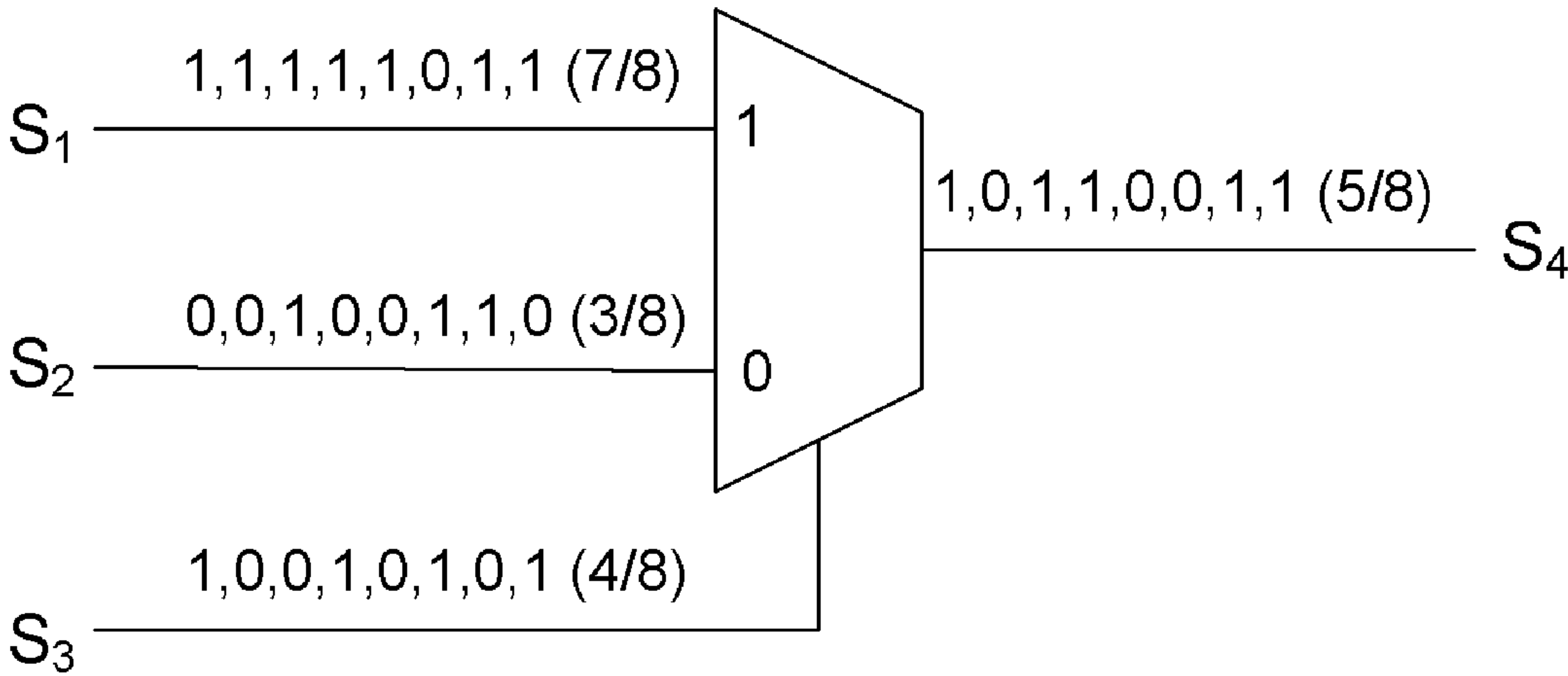
【圖6F】



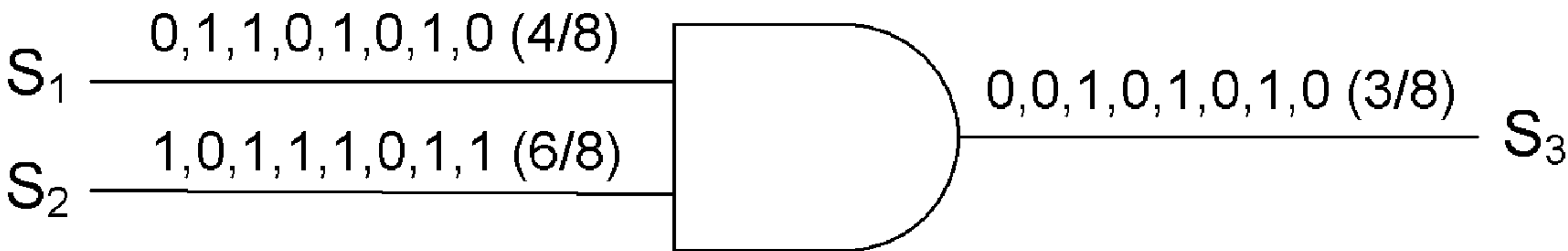
【圖6G】



【圖7】

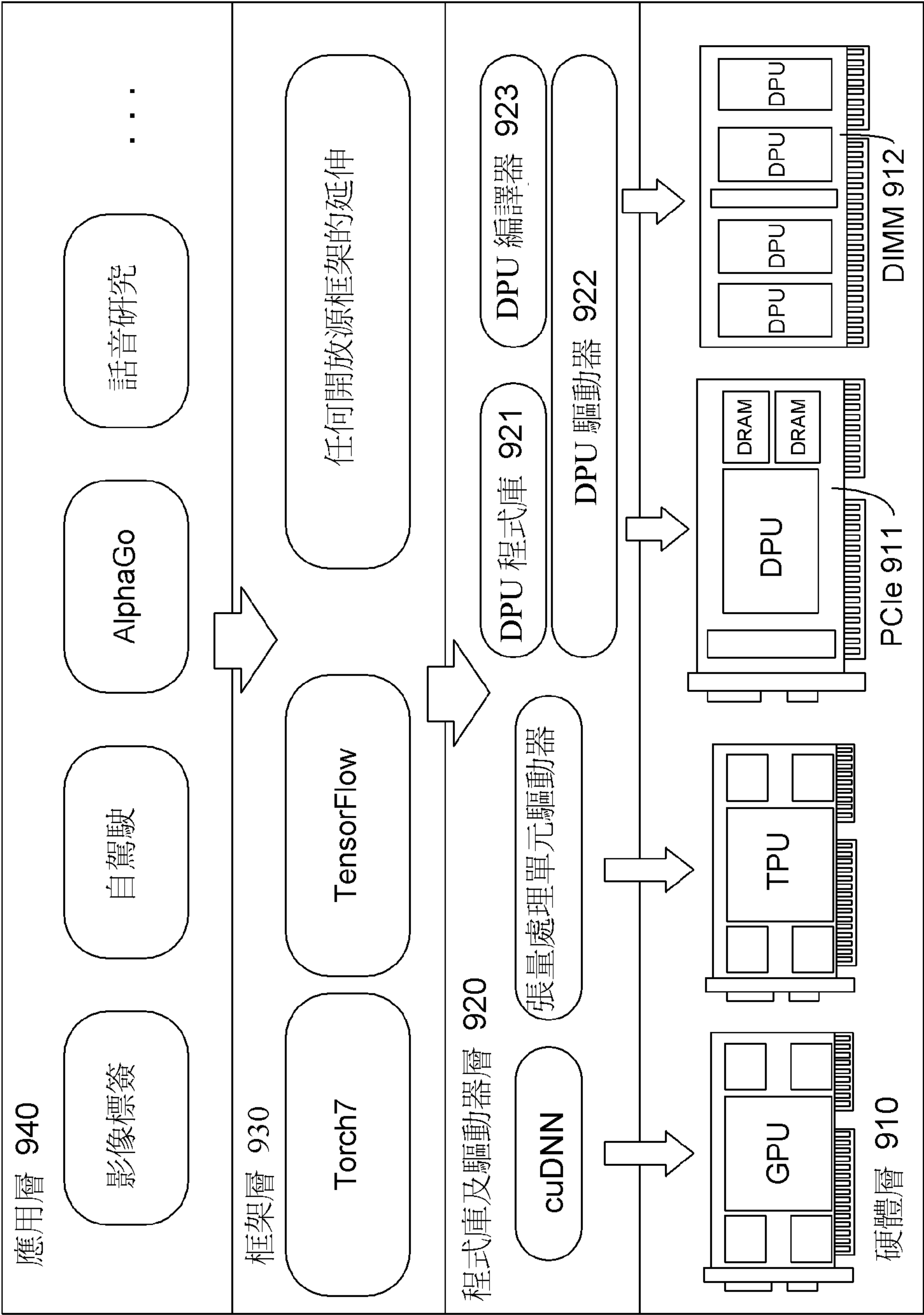


【圖8A】



【圖8B】

900



【圖9】