



(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 I848184 B

(45)公告日：中華民國 113 (2024) 年 07 月 11 日

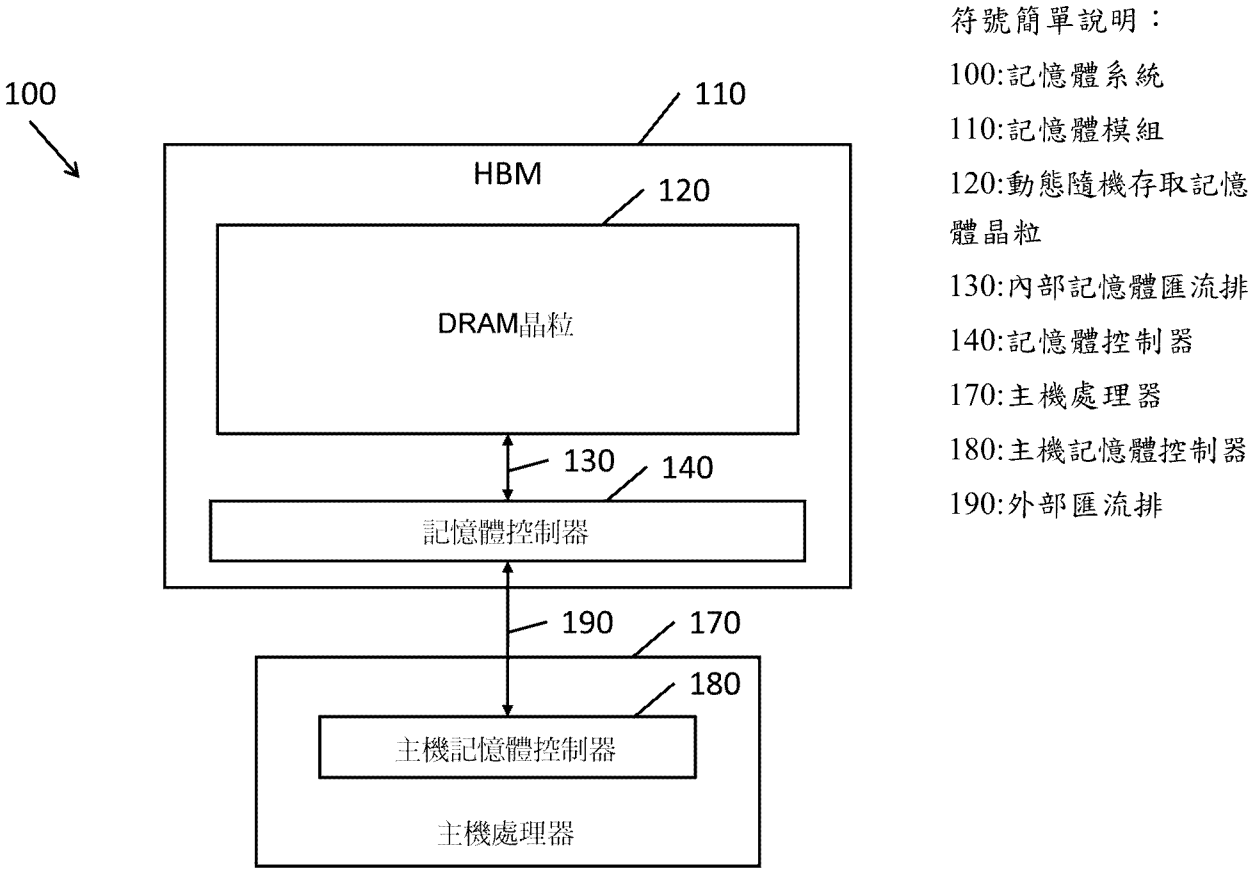
(21)申請案號：109138406 (22)申請日：中華民國 109 (2020) 年 11 月 04 日
(51)Int. Cl. : G06F12/02 (2006.01) G06F12/0802(2016.01)
(30)優先權：2020/02/12 美國 62/975,577
2020/04/27 美國 16/859,829
(71)申請人：南韓商三星電子股份有限公司 (南韓) SAMSUNG ELECTRONICS CO., LTD. (KR)
南韓
(72)發明人：馬拉迪 克里希納 特佳 MALLADI, KRISHNA TEJA (IN) ; 皇甫文沁 HUANGFU,
WENQIN (CN)
(74)代理人：林孟閱；盧珮君；陳怡如
(56)參考文獻：
TW 201816785A CN 109669887A
US 6295586B1 US 10185499B1
審查人員：陳延慶
申請專利範圍項數：24 項 圖式數：10 共 67 頁

(54)名稱
記憶體模組以及用於在記憶體內進行計算的方法
(57)摘要

本發明提供一種記憶體模組，包含：記憶體晶粒，包含動態隨機存取記憶（DRAM）庫，每一 DRAM 庫包含：DRAM 胞陣列，以頁配置；列緩衝器，用以儲存頁中的一者的值；輸入/輸出（IO）模組；以及記憶體內計算（IMC）模組，包含：算術邏輯單元（ALU），用以自列緩衝器或 IO 模組接收運算元，以及基於運算元及多個 ALU 運算中的一者來計算輸出；以及結果暫存器，用以儲存 ALU 的輸出；以及控制器，用以：自主機處理器接收運算元及指令；基於指令來判定資料佈局；根據資料佈局將運算元供應至 DRAM 庫；以及控制 IMC 模組以根據指令對運算元進行 ALU 運算中的一者。

A memory module includes: a memory die including a dynamic random access memory (DRAM) banks, each including: an array of DRAM cells arranged in pages; a row buffer to store values of one of the pages; an input/output (IO) module; and an in-memory compute (IMC) module including: an arithmetic logic unit (ALU) to receive operands from the row buffer or the IO module and to compute an output based on the operands and one of a plurality of ALU operations; and a result register to store the output of the ALU; and a controller to: receive, from a host processor, operands and an instruction; determine, based on the instruction, a data layout; supply the operands to the DRAM banks in accordance with the data layout; and control an IMC module to perform one of the ALU operations on the operands in accordance with the instruction.

指定代表圖：



【圖1】



I848184

【發明摘要】

【中文發明名稱】記憶體模組以及用於在記憶體內進行計算的方法

【英文發明名稱】 MEMORY MODULE AND METHOD FOR PERFORMING COMPUTATIONS IN-MEMORY

【中文】本發明提供一種記憶體模組，包含：記憶體晶粒，包含動態隨機存取記憶（DRAM）庫，每一DRAM庫包含：DRAM胞陣列，以頁配置；列緩衝器，用以儲存頁中的一者的值；輸入/輸出（IO）模組；以及記憶體內計算（IMC）模組，包含：算術邏輯單元（ALU），用以自列緩衝器或IO模組接收運算元，以及基於運算元及多個ALU運算中的一者來計算輸出；以及結果暫存器，用以儲存ALU的輸出；以及控制器，用以：自主機處理器接收運算元及指令；基於指令來判定資料佈局；根據資料佈局將運算元供應至DRAM庫；以及控制IMC模組以根據指令對運算元進行ALU運算中的一者。

【英文】A memory module includes: a memory die including a dynamic random access memory (DRAM) banks, each including: an array of DRAM cells arranged in pages; a row buffer to store values of one of the pages; an input/output (IO) module; and an in-memory compute (IMC) module including: an arithmetic logic unit (ALU) to receive operands from the row buffer or the IO module and to compute an output based on the operands and one of a plurality of ALU operations; and a result register to store the output of the ALU; and a controller to: receive, from a host processor, operands and an instruction; determine, based

on the instruction, a data layout; supply the operands to the DRAM banks in accordance with the data layout; and control an IMC module to perform one of the ALU operations on the operands in accordance with the instruction.

【指定代表圖】圖 1

【代表圖之符號簡單說明】

100:記憶體系統

110:記憶體模組

120:動態隨機存取記憶體晶粒

130:內部記憶體匯流排

140:記憶體控制器

170:主機處理器

180:主機記憶體控制器

190:外部匯流排

【特徵化學式】

無

【發明說明書】

【中文發明名稱】記憶體模組以及用於在記憶體內進行計算的方法

【英文發明名稱】 MEMORY MODULE AND METHOD FOR PERFORMING COMPUTATIONS IN-MEMORY

【技術領域】

【0001】 本揭露內容的實施例的態樣是關於用於記憶體內計算的資料置放的系統及方法。

【先前技術】

【0002】 高頻寬記憶體（High Bandwidth Memory；HBM）通常用作圖形處理單元（graphics processing unit；GPU）的高效能記憶體。HBM 具有相較於典型 DRAM 而具有極寬匯流排的優點。當前 HBM 架構包含可使用矽穿孔（through silicon via；TSV）來連接的多個堆疊的 DRAM 晶粒（例如，小方塊（dice）），以及充當 HBM 中的緩衝器及 GPU 中的 HBM 記憶體控制器的邏輯晶粒。可藉由將記憶體中的過程（例如，記憶體內處理）能力添加至記憶體系統來達成進一步效能增益。

【0003】 上述資訊僅用於增進對本揭露內容的實施例的背景的理解，且因此可能含有未構成先前技術的資訊。

【發明內容】

【0004】 根據本揭露內容的一個實施例，一種記憶體模組包含：記

記憶體晶粒，包含多個動態隨機存取記憶（dynamic random access memory；DRAM）庫，所述 DRAM 庫中的每一者包含：DRAM 胞陣列，以多個頁配置，每一頁包含多個 DRAM 胞，每一胞儲存位元值；列緩衝器，經組態以儲存所述多個頁中的打開頁的值；輸入/輸出（input/output；IO）模組；以及記憶體內計算（in-memory compute；IMC）模組，包含：算術邏輯單元（arithmetic logic unit；ALU），經組態以自所述列緩衝器或所述 IO 模組接收運算元，以及基於所述運算元及多個 ALU 運算中的所選 ALU 運算來計算輸出；以及結果暫存器，經組態以儲存由所述 ALU 計算出的所述輸出；以及記憶體控制器，經組態以：自主機處理器接收第一運算元、第二運算元以及指令；基於所述指令自多個資料佈局判定資料佈局；根據所述資料佈局將所述第一運算元及所述第二運算元供應至所述 DRAM 庫；以及控制所述 DRAM 庫的 IMC 模組以根據所述指令對所述第一運算元及所述第二運算元進行所述多個 ALU 運算中的 ALU 運算。

【0005】 所述資料佈局可包含：一個運算元（one operand；1OP）資料佈局，其中將所述第一運算元寫入至所述 DRAM 胞且將所述第二運算元自所述主機處理器直接供應至所述 DRAM 庫的所述 IMC 模組。

【0006】 所述 IMC 模組可更包含運算元暫存器，且所述記憶體控制器可進一步經組態以：將所述第一運算元的第一磚片儲存於所述運算元暫存器中；以及對儲存於所述運算元暫存器中的所述第一運算元及所述第二運算元的多個第二磚片中的每一者進行所述 ALU 運算。

【0007】 可將所述第一運算元劃分成多個第一磚片且將所述第二運算元劃分成多個第二磚片，每一磚片包含多個值，且其中所述資料佈局可包含相同頁（same page；SP）資料佈局，其中所述記憶體控制器將所述第一磚片中的一或多者及所述第二磚片中的一或多者儲存於所述 DRAM 胞的相同頁中。

【0008】 所述 IMC 模組可更包含運算元暫存器，且所述記憶體控制器可進一步經組態以：將所述第一磚片中的一或多者中的第一磚片儲存於所述運算元暫存器中；以及對儲存於所述運算元暫存器中的所述第一運算元及與所述第一磚片儲存於所述 DRAM 胞陣列的所述相同頁中的所述一或多個第二磚片中的每一者進行所述 ALU 運算。

【0009】 至少一個 DRAM 庫的所述 IMC 模組可更包含累加器，所述累加器包含經組態以儲存經累加值的累加器暫存器，所述累加器經組態以：接收由所述 ALU 計算出的所述輸出；以及用所述經累加值與所述輸出的總和更新所述累加器暫存器，且所述指令可包含計算所述第一運算元與所述第二運算元的內積，其中所述第一磚片中的所述第一磚片儲存列資料且所述第二磚片中的第二磚片包含行資料。

【0010】 所述第一磚片可具有第一數目個值且所述第二磚片具有第二數目個值，至少一個 DRAM 庫的所述 IMC 模組可包含輸出緩衝器，所述輸出緩衝器具有用於儲存大於或等於所述第一數目個值與所述第二數目個值的乘積的大小，且所述指令可包含計算所述第一運算元與所述第二運算元的外積，其中所述第一磚片中的所述第一磚片儲存列資料且所述第二磚片中的第二磚片包含行資

料。

【0011】 所述第一磚片可具有第一數目個值且所述第二磚片可具有第二數目個值，至少一個 DRAM 庫的所述 IMC 模組可包含輸出緩衝器，所述輸出緩衝器具有用於儲存大於或等於所述第一數目個值及所述第二數目個值中的更大者的大小，且所述指令可包含計算所述第一運算元與所述第二運算元的張量積，其中所述第一磚片中的所述第一磚片儲存列資料且所述第二磚片中的第二磚片包含行資料。

【0012】 可將所述第一運算元劃分成多個第一磚片且可將所述第二運算元劃分成多個第二磚片，每一磚片包含多個值，且所述資料佈局可包含不同頁（different page；DP）資料佈局，其中所述記憶體控制器儲存所述 DRAM 胞陣列的第一頁中的所述第一磚片的子集及所述 DRAM 胞陣列的第二頁中的所述第二磚片的子集。

【0013】 所述 IMC 模組可更包含運算元暫存器，且所述記憶體控制器可進一步經組態以：將來自所述第一頁的所述第一運算元的第一磚片儲存於所述運算元暫存器中；以及對儲存於所述運算元暫存器中的所述第一運算元及來自所述第二頁的所述第二運算元的多個第二磚片中的每一者進行所述 ALU 運算。

【0014】 每一 DRAM 庫的所述 IMC 模組可更包含經組態以緩衝由所述 ALU 計算出的所述輸出的硬體緩衝器。

【0015】 所述硬體緩衝器可為所述結果暫存器的大小的至少四倍。

【0016】 每一 DRAM 庫的所述 IMC 模組可更包含累加器，所述累加器包含經組態以儲存經累加值的累加器暫存器，所述累加器經

組態以：接收由所述 ALU 計算出的所述輸出；以及用所述經累加值與所述輸出的總和更新所述累加器暫存器。

【0017】 所述記憶體模組可為包含藉由矽穿孔連接的記憶體晶粒堆疊的高頻寬記憶體（HBM）模組，所述多個記憶體晶粒包含所述記憶體晶粒。

【0018】 根據本揭露內容的一個實施例，一種用於在記憶體內進行計算的方法包含：由記憶體模組的記憶體控制器接收第一運算元、第二運算元、指令；由所述記憶體控制器基於所述指令自多個資料佈局判定資料佈局；根據所述資料佈局將所述第一運算元及所述第二運算元供應至所述記憶體模組的一或多個動態隨機存取記憶體（DRAM）庫，所述 DRAM 庫中的每一者包含：DRAM 胞陣列，以多個頁配置，每一頁包含多個 DRAM 胞，每一胞儲存位元值；列緩衝器，經組態以儲存所述多個頁中的打開頁的值；輸入/輸出（IO）模組；以及記憶體內計算（IMC）模組，包含：算術邏輯單元（ALU），經組態以自所述列緩衝器或所述 IO 模組接收運算元，以及基於所述運算元及多個 ALU 運算中的所選 ALU 運算來計算輸出；以及結果暫存器，經組態以儲存由所述 ALU 計算出的所述輸出；以及控制所述 DRAM 庫的所述 IMC 模組以根據所述指令對所述第一運算元及所述第二運算元進行所述多個 ALU 運算中的 ALU 運算。

【0019】 所述資料佈局可包含：一個運算元（1OP）資料佈局，其中將所述第一運算元寫入至所述 DRAM 胞且將所述第二運算元自所述主機處理器直接供應至所述 DRAM 庫的所述 IMC 模組。

【0020】 所述 IMC 模組可更包含運算元暫存器，且所述記憶體控

制器可進一步經組態以：將所述第一運算元的第一磚片儲存於所述運算元暫存器中；以及對儲存於所述運算元暫存器中的所述第一運算元及所述第二運算元的多個第二磚片中的每一者進行所述 ALU 運算。

【0021】 可將所述第一運算元劃分成多個第一磚片且將所述第二運算元劃分成多個第二磚片，每一磚片包含多個值，且所述資料佈局可包含相同頁（SP）資料佈局，其中所述記憶體控制器將所述第一磚片中的一或多者及所述第二磚片中的一或多者儲存於所述 DRAM 胞的相同頁中。

【0022】 所述 IMC 模組可更包含運算元暫存器，且所述記憶體控制器可進一步經組態以：將所述第一磚片中的一或多者中的第一磚片儲存於所述運算元暫存器中；以及對儲存於所述運算元暫存器中的所述第一運算元及與所述第一磚片儲存於所述 DRAM 胞陣列的所述相同頁中的所述一或多個第二磚片中的每一者進行所述 ALU 運算。

【0023】 至少一個 DRAM 庫的所述 IMC 模組可更包含累加器，所述累加器包含經組態以儲存經累加值的累加器暫存器，所述累加器經組態以：接收由所述 ALU 計算出的所述輸出；以及用所述經累加值與所述輸出的總和更新所述累加器暫存器，且其中所述指令包含計算所述第一運算元與所述第二運算元的內積，其中所述第一磚片中的所述第一磚片儲存列資料且所述第二磚片中的第二磚片包含行資料。

【0024】 所述第一磚片可具有第一數目個值且所述第二磚片可具有第二數目個值，至少一個 DRAM 庫的所述 IMC 模組可包含輸出

緩衝器，所述輸出緩衝器具有用於儲存大於或等於所述第一數目個值與所述第二數目個值的乘積的大小，且所述指令可包含計算所述第一運算元與所述第二運算元的外積，其中所述第一磚片中的所述第一磚片儲存列資料且所述第二磚片中的第二磚片包含行資料。

【0025】 所述第一磚片可具有第一數目個值且所述第二磚片可具有第二數目個值，至少一個 DRAM 庫的所述 IMC 模組可包含輸出緩衝器，所述輸出緩衝器具有用於儲存大於或等於所述第一數目個值及所述第二數目個值中的更大者的大小，且所述指令可包含計算所述第一運算元與所述第二運算元的張量積，其中所述第一磚片中的所述第一磚片儲存列資料且所述第二磚片中的第二磚片包含行資料。

【0026】 可將所述第一運算元劃分成多個第一磚片且將所述第二運算元劃分成多個第二磚片，每一磚片包含多個值，且所述資料佈局可包含不同頁（DP）資料佈局，其中所述記憶體控制器儲存所述 DRAM 胞陣列的第一頁中的所述第一磚片的子集及所述 DRAM 胞陣列的第二頁中的所述第二磚片的子集。

【0027】 所述 IMC 模組可更包含運算元暫存器，且所述記憶體控制器可進一步經組態以：將來自所述第一頁的所述第一運算元的第一磚片儲存於所述運算元暫存器中；以及對儲存於所述運算元暫存器中的所述第一運算元及來自所述第二頁的所述第二運算元的多個第二磚片中的每一者進行所述 ALU 運算。

【圖式簡單說明】

【0028】

圖 1 為根據本揭露內容的實施例的高頻寬記憶體（HBM）系統的架構的方塊圖。

圖 2A 為根據本揭露內容的一個實施例的具有嵌入式算術邏輯單元（ALU）的記憶體的庫的示意性方塊圖。

圖 2B 為根據本揭露內容的一個實施例的示出 DRAM 胞的實施例的電路圖。

圖 2C 為根據本揭露內容的一個實施例的 DRAM 塊的示意圖。

圖 3 為根據本揭露內容的一個實施例的 DRAM 塊陣列的示意圖。

圖 4A 示意性地示出根據本揭露內容的一個實施例的在一個運算元（矩陣 A）儲存於 DRAM 中且第二運算元（矩陣 B）自具有積體記憶體內計算（IMC）的記憶體模組的外部廣播的情況下用於通用矩陣-矩陣乘法（general matrix-matrix multiplication；GEMM）的資料的置放。

圖 4B 示意性地示出根據本揭露內容的一個實施例的在兩個運算元（矩陣 A 及矩陣 B）儲存於具有積體 IMC 的記憶體模組的 DRAM 的相同頁中的情況下用於通用矩陣-矩陣乘法（GEMM）的資料的置放。

圖 4C 示意性地示出根據本揭露內容的一個實施例的在兩個運算元（矩陣 A 及矩陣 B）儲存於具有積體 IMC 的記憶體模組的 DRAM 的不同頁中的情況下用於通用矩陣-矩陣乘法（GEMM）的資料的置放。

圖 5A 為根據本揭露內容的一個實施例的一個運算元（1OP）資料佈局中的矩陣 A 的第一列乘以矩陣 B 的第一行的乘法的示意性描繪，其中一個運算元自外部饋送且一個運算元儲存於具有積體 IMC 的 DRAM 庫中。

圖 5B 為根據本揭露內容的一個實施例的 1OP 資料佈局中的矩陣 A 的第一列的第一值乘以矩陣 B 的每一行的第一值的乘法（其中資料再用）的示意性描繪，其中一個運算元自外部饋送且一個運算元儲存於具有積體 IMC 的 DRAM 庫中。

圖 6A 為根據本揭露內容的一個實施例的相同列（same row；SR）資料佈局中的矩陣 A 的第一列乘以矩陣 B 的第一行的乘法的示意性描繪，其中兩個運算元儲存於具有積體 IMC 的 DRAM 庫的相同頁中。

圖 6B 為根據本揭露內容的一個實施例的 SR 資料佈局中的矩陣 A 的第一列的第一值乘以矩陣 B 的每一行的第一值的乘法（其中資料再用）的示意性描繪，其中兩個運算元儲存於具有積體 IMC 的 DRAM 庫的相同頁中。

圖 7A 為根據本揭露內容的一個實施例的不同列（different row；DR）資料佈局中的矩陣 A 的第一列乘以矩陣 B 的第一行的乘法的示意性描繪，其中運算元儲存於具有記憶體內計算的 DRAM 庫的不同頁中。

圖 7B 為根據本揭露內容的一個實施例的 DR 資料佈局中的矩陣 A 的第一列的第一值乘以矩陣 B 的每一行的第一值及矩陣 A 的第一列的第二值乘以矩陣 B 的每一行的第二值的乘法（其中資料再用）的示意性描繪，其中運算元儲存於具有記憶體內計算的

DRAM 庫的不同頁中。

圖 8 為根據本揭露內容的一個實施例的 DRAM 庫的記憶體內計算（IMC）模組的示意性方塊圖，其中 IMC 模組進一步連接至結果緩衝器、累加器以及緩衝器。

圖 9 為根據本揭露內容的一些實施例的用以解釋使用相同列（SR）佈局的計算的通用矩陣-矩陣乘法（GEMM）的實施例的示意性圖示。

圖 10 為根據本揭露內容的實施例的用於控制資料在具有記憶體內計算模組的 DRAM 庫中的置放的方法的流程圖。

【實施方式】

【0029】 參考以下實施例的具體描述及隨附圖式可更易於理解所揭露概念以及實現所揭露概念的方法的特徵。在下文中，將參考隨附圖式更詳細地描述實施例，其中相同附圖標號始終指代相同元件。然而，本揭露內容可以各種不同形式體現，且不應解釋為僅限於本文中的所示出實施例。確切而言，提供此等實施例作為實例，使得本揭露內容將為透徹且完整的，且將向所屬技術領域中具有通常知識者充分傳達本揭露內容的態樣及特徵。因此，可能未描述不為所屬技術領域中具有通常知識者完整理解本揭露內容的態樣及特徵所必需的過程、元件以及技術。除非另外指出，否則相同附圖標號貫穿隨附圖式及書面描述標示相同元件，且因此將不重複其描述。在圖式中，為了清楚起見，可放大元件、層以及區的相對大小。

【0030】 在以下描述中，出於解釋的目的，闡述眾多具體細節以提

供對各種實施例的透徹理解。然而，顯而易見的是，可在不具有此等特定細節或具有一或多個等效配置的情況下實踐各種實施例。在其他情況下，以方塊圖繪示熟知的結構及裝置，以便避免不必要地混淆各種實施例。

【0031】本文中所使用的術語僅出於描述特定實施例的目的，且並不意欲限制本揭露內容。如本文中所使用，除非上下文另外明確指示，否則單數形式「一 (a/an)」亦意欲包含複數形式。應進一步理解，術語「包括 (comprises/comprising)」、「具有 (have/having)」以及「包含 (includes/including)」在用於本說明書中時指定所陳述特徵、整數、步驟、操作、元件及/或組件的存在，但不排除一或多個其他特徵、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「及/或」包含相關聯的所列項目中的一或多者的任何以及所有組合。

【0032】本文中所描述的根據本揭露內容的實施例的電子裝置或電氣裝置及/或任何其他相關裝置或組件可利用任何合適的硬體、韌體（例如，特殊應用積體電路）、軟體，或軟體、韌體以及硬體的組合來實施。舉例而言，在一些情況下，此等裝置的各種組件可形成於一個積體電路（integrated circuit；IC）晶片上或單獨的 IC 晶片上。另外，此等裝置的各種組件可實施於可撓性印刷電路膜、帶載體封裝 (tape carrier package；TCP)、印刷電路板 (printed circuit board；PCB) 上或形成於一個基底上。另外，此等裝置的各種組件可為過程或線程，其在一或多個處理器上運行、在一或多個計算裝置中運行、執行電腦程式指令且與其他系統組件交互以用於進行本文中所描述的各種功能性。電腦程式指令儲存於記憶體中，所述

記憶體可使用諸如（例如）隨機存取記憶體（random access memory；RAM）的標準記憶體裝置來實施於計算裝置中。電腦程式指令亦可儲存於諸如（例如）CD-ROM、快閃驅動器或類似者的其他非暫時性電腦可讀媒體中。此外，所屬技術領域中具有通常知識者應認識到，在不脫離本揭露內容的例示性實施例的精神及範圍的情況下，各種計算裝置的功能性可經組合或整合至單一計算裝置中，或特定計算裝置的功能性可跨一或多個其他計算裝置分佈。

【0033】 除非另外定義，否則本文中所使用的所有術語（包含技術及科學術語）具有與本揭露內容所屬技術領域中具有通常知識者通常理解的相同意義。應進一步理解，術語（諸如，常用詞典中所定義的彼等術語）應解釋為具有與其在相關技術及/或本說明書的上下文中的含義一致的含義，且不應以理想化或過分正式意義來進行解釋，除非本文中明確地如此定義。

【0034】 本揭露內容的一些態樣的實施例大體上是關於在記憶體內計算的上下文中管理資料在記憶體中的置放。記憶體內計算的上下文的一個實例是在高頻寬記憶體（HBM）系統中，所述 HBM 系統包含含有記憶體的動態隨機存取記憶體（DRAM）晶粒以及 HBM 邏輯晶粒上的算術邏輯單元（ALU）及記憶體控制器，其中 HBM 邏輯晶粒上的 ALU 可進行記憶體內計算。HBM 的記憶體控制器控制將資料儲存至 DRAM 晶粒的記憶體中及自 DRAM 晶粒讀取資料。

【0035】 出於清楚起見，如本文中所使用，術語「記憶體內計算」是指在不遍歷外部資料匯流排的情況下使用儲存於 DRAM 晶粒中

的資料在諸如高頻寬記憶體模組的記憶體模組內進行計算。在比較電腦系統中，處理器可經由外部 DRAM 資料匯流排耦接至主記憶體（例如，DRAM），且自主記憶體存取資料比存取處理器內的暫存器檔案中的資料及/或更靠近處理器的硬體快取（例如，L1 快取、L2 快取）中的資料顯著更慢（例如，慢幾個數量級）。藉由將額外處理器置放於記憶體處或記憶體附近（例如，「記憶體內處理器」），可避免因遍歷外部匯流排所致的潛時，由此得到更高效能計算。

【0036】 本揭露內容的實施例的態樣是關於將諸如 ALU 的計算電路系統置放於與 DRAM 庫相同的晶粒上，例如直接連接至各別 DRAM 庫的感測放大器或列緩衝器。

【0037】 歸因於 DRAM 模組的設計及效能特性，記憶體中的資料的特定配置可影響記憶體內計算的效能。因此，本揭露內容的實施例的一些態樣是關於用於將資料置放於記憶體模組（例如，HBM 記憶體模組）的 DRAM 模組內的系統及方法，其中資料的置放可基於待由記憶體內計算（IMC）模組執行的計算的特性來控制。

【0038】 圖 1 為根據本揭露內容的實施例的記憶體（例如，HBM）系統的架構的方塊圖。

【0039】 參考圖 1，本揭露內容的實施例提供用於記憶體內函數（function-in-memory；FIM）記憶體系統 100 的系統。記憶體系統 100（或 HBM 系統）支援待整合於記憶體模組（或 HBM 模組）110 中的額外計算資源。舉例而言，在各種實施例中，記憶體系統 100 允許待在記憶體內進行的一些資料計算及移動，且亦提供高容量記憶體草稿。記憶體系統 100 包含連接至主機處理器 170 的至

少一個記憶體模組 110，所述主機處理器 170 諸如圖形處理單元（GPU）或中央處理單元（central processing unit；CPU）。在各種實施例中，記憶體模組 110 包含經由內部記憶體匯流排 130 連接至（例如，邏輯晶粒上的）記憶體控制器 140 的一或多個動態隨機存取記憶體（DRAM）晶粒 120。在各種實施例中，主機處理器 170 包含用於與記憶體模組 110 介接的主機記憶體控制器（或主機控制器）180。然而，本揭露內容的實施例不限於此。舉例而言，主機記憶體控制器 180 可與主機處理器 170 分開（例如，在單獨晶粒上或與主機處理器 170 在相同晶粒上）。

【0040】根據各種實施例，記憶體控制器 140 經組態以協調對來自主機處理器 170 的指令的執行。指令可包含常規指令及 FIM 指令兩者。舉例而言，常規指令（例如，傳統載入及儲存函數，不是記憶體內函數指令）由主機記憶體控制器 180 發送且由記憶體控制器 140 接收並以習知方式執行。舉例而言，常規指令可包含將在外匯流排 190 上接收到的資料儲存於 DRAM 晶粒 120 中的指令，以及自 DRAM 晶粒 120 檢索資料且將所述資料經由外部匯流排 190 傳送至主機處理器 170 的指令。在一些實施例中，常規指令及 FIM 指令可包含用於將資料儲存於 DRAM 晶粒中的特定方位（例如，特定庫的特定頁）中的操作。此等資料可包含兩個不同運算元，其中每一運算元可包含多個值（例如，浮點或整數值），且如下文更詳細地論述，此等運算元的值可根據各種資料置放策略跨 DRAM 晶粒的不同方位分佈。

【0041】本揭露內容的實施例的態樣是關於記憶體內計算（IMC）的使用。一些比較 HBM 系統包含位於 DRAM 晶粒 120 外部（例

如，位於記憶體控制器 140 處）且由 DRAM 晶粒 120 的記憶體庫共用的算術邏輯單元（ALU），使得 ALU 可在不遍歷外部匯流排 190 的情況下對儲存於一或多個 DRAM 晶粒 120 中的資料進行運算（例如，算術運算）。舉例而言，記憶體控制器 140 可執行資料移動操作（例如，載入/儲存對指令）以在 DRAM 晶粒 120 的不同部分之間移動或複製資料。舉例而言，記憶體控制器 140 可諸如藉由協調對利用 ALU 的計算 FIM 指令（例如，基元指令及 ALU 指令）的執行來執行最初為多個常規指令的 FIM 指令。作為另一實例，在一些情況下，自主機處理器 170 接收到的 FIM 指令使得具有記憶體 IMC 的記憶體模組基於所供應輸入使用經訓練機器學習模型（例如，神經網路）來進行特定計算（諸如計算推理）、在訓練神經網路期間進行反向傳播，或將兩個矩陣相乘。在此等情況下，記憶體控制器 140 可藉由將接收到的資料（例如，指令的運算元）儲存於 DRAM 庫的特定頁中，且藉由將特定 ALU 運算提供至儲存與 FIM 指令相關的資料的 DRAM 庫來協調對此等指令的執行。記憶體控制器 140 可在執行特定 FIM 指令時以改良計算的效能的方式將接收到的資料（運算元）置放於 DRAM 庫的特定頁中。接著將結果儲存於 DRAM 晶粒 120 中或可將結果經由外部匯流排 190 返回至主機處理器 170。

【0042】 本揭露內容的實施例的一些態樣是關於藉由將記憶體內計算（IMC）模組整合至 DRAM 晶粒 120 的記憶體庫中來加速記憶體有界運算，由此避免 DRAM 晶粒 120 與記憶體控制器 140 之間的內部記憶體匯流排 130 的遍歷。舉例而言，IMC 模組與保存資料的 DRAM 庫位於相同的實體半導體晶粒（physical

semiconductor) 上。每一 DRAM 庫可具有相關聯 IMC 模組，使得可在不遍歷內部記憶體匯流排 130 至記憶體控制器的情況下（例如，在不將資料傳送出 DRAM 庫的情況下）對儲存於 DRAM 庫中的資料進行計算，同時亦使得計算能夠跨 DRAM 庫的 IMC 模組並行化。

【0043】 儘管本揭露內容的實施例的態樣將在高頻寬記憶體的上下文中進行描述，但實施例不限於此，且亦可適用於將 IMC 模組整合至其他類型的 DRAM 系統中的 DRAM 晶粒中。

【0044】 圖 2A 為根據本揭露內容的一個實施例的具有嵌入式算術邏輯單元 (ALU) 的記憶體的庫的示意性方塊圖。如圖 2A 中所繪示，DRAM 庫 200 包含配置成列及行（或頁及行）的 DRAM 胞 210 陣列。在圖 2A 中所繪示的實例中，DRAM 庫 200 包含配置成 n 個列（或頁）及 m 個行的 DRAM 胞 210。多個位元線 B1 至位元線 B m 沿行方向延伸，且多個列賦能線 R1 至列賦能線 R n 沿陣列的列方向延伸且與位元線交叉。位元線中的每一者連接至其對應行中的每一胞（例如，陣列的第 i 行中的每一胞連接至位元線 B i ）。同樣，列賦能線 R1 至列賦能線 R n 中的每一者連接至其對應列中的每一 DRAM 胞 210（例如，陣列的第 j 列或第 j 頁中的每一胞連接至列賦能線 R j ）。DRAM 庫 200 的 DRAM 胞 210 的列亦可稱為 DRAM 頁。

【0045】 圖 2B 為根據本揭露內容的一個實施例的示出 DRAM 胞的實例的電路圖。每一 DRAM 胞 210 可大體上經模型化為包含用於儲存資料電壓（例如，位元值，其中每一電容器可儲存表示 0 位元的電壓或表示 1 位元的電壓）的電容器 212 及用於將資料電壓

傳輸至電容器 212 的開關 214。圖 2B 中所繪示的特定 DRAM 胞 210 位於陣列的第 j 列及第 i 行中。因此，圖 2B 的 DRAM 胞 210 的開關 214 連接於第 i 位元線 B_i 與電容器 212 的一個端之間，且電容器 212 的另一端連接至接地。圖 2B 中所繪示的 DRAM 胞 210 的開關 214 的閘電極連接至第 j 列賦能線 R_j ，使得當開關 214 接通時電容器 212 連接至位元線 B_i 。

【0046】 返回參考圖 2A，DRAM 庫 200 包含連接至列賦能線 R_1 至列賦能線 R_n 的列解碼器 220，且列解碼器 220 經組態以將列賦能信號供應至對應於自例如記憶體控制器 140 供應的列位址的列賦能線中的特定一者。當將資料寫入或讀取至 DRAM 胞的特定列（或頁）時，列解碼器 220 將列賦能信號供應至對應於特定列（或頁）的列賦能線。當寫入資料時，在啟用特定列（或頁）的同時，將對應於待寫入的資料的電壓供應至位元線 B_1 至位元線 B_m 。類似地，當自 DRAM 胞 210 陣列的特定列（或頁）讀取資料時，對應於儲存於電容器 212 中的電壓的電壓沿位元線 B_1 至位元線 B_m 傳輸，且由包含感測放大器 232 的輸入/輸出感測放大器層 230（或 input/output sense amplifier；IOSA）讀取。感測放大器 232 中的每一感測放大器連接至位元線中的對應一者（例如，感測放大器 232 包含 m 個感測放大器）。舉例而言，在一些實施例中，DRAM 胞 210 陣列包含 8,192 個行，以及連接至 8,192 個對應感測放大器 232 的 8,192 個對應位元線（例如，位元線 B_1 至位元線 B_{8192} ）（例如，每一頁可儲存 8,192 個位元或 8 千位元的資料）。由於感測放大器 232 儲存自當前列（或頁）讀取的資料直至所述資料藉由「預充電」命令清除為止，故感測放大器 232 亦可稱為「列緩衝器」。

【0047】 行解碼器 240 可用於使用多工器 234 來選擇資料行的子集，且可通過經由全域 IO 層 236 將所讀取的資料供應至記憶體內計算（IMC）模組 250 以供對資料進行計算。舉例而言，在一些實施例中，行解碼器 240 及多工器 234 可允許自 DRAM 胞 210 的 8,192 個行選擇 256 個位元（256b）的資料。

【0048】 當自 DRAM 庫 200 的不同於當前儲存於感測放大器 232 中的頁的頁載入資料時，使用「預充電」（PRE）命令來關閉當前頁以準備 DRAM 庫 200 以供下一存取。接著使用「激活」（ACT）命令來打開庫的特定列或頁且接著將資料儲存於感測放大器 232 中。接著可自感測放大器 232 讀取（READ）資料以傳送至 IMC 模組 250。

【0049】 另一方面，當將資料自己打開的頁載入至 IMC 模組 250 中時，則 READ 命令可足以諸如藉由使用行解碼器 240 選擇已儲存於感測放大器 232 中的資料的適當子集來載入資料（可跳過 PRE 命令及 ACT 命令）。

【0050】 根據各種實施例，IMC 模組 250（或 ALU&Reg）包含算術邏輯單元（ALU）252 及一或多個暫存器。在圖 2A 中所繪示的實施例中，IMC 模組 250 包含運算元暫存器 Rop 254（或輸入緩衝器）及結果暫存器 Rz 256。多工器 257 及多工器 258 可用於控制去往 ALU 252 的兩個輸入的資料流（例如，作為去往 ALU 252 的第一運算元及第二運算元）。舉例而言，在圖 2A 中所繪示的實施例中，運算元暫存器 Rop 254 連接至 ALU 252 的第一運算元輸入，且連接第一多工器 257 以將資料自感測放大器 232 經由全域 IO 層 236 或自外部源（例如，主機處理器）經由輸入-輸出（IO）模組

260（或寫入輸入/輸出及讀取輸入/輸出，或 write input/output；WIO 及 read input/output；RIO）寫入至運算元暫存器 Rop 254。在圖 2A 中所繪示的實施例中，第二多工器 258 經組態以將資料再次自感測放大器 232 經由全域 IO 層 236 或自外部源經由 IO 模組 260 直接供應至 ALU 252 的第二運算元輸入。ALU 252 可將其計算輸出至結果暫存器 Rz 256，且資料可自結果暫存器 Rz 256 經由全域 IO 層 236 寫回至 DRAM 胞 210 中，或經由 IO 模組 260（或 WIO 及 RIO）傳輸出至主機處理器 170。

【0051】根據一些實施例，ALU 252 經組態以用於執行各種計算運算（例如，簡單計算命令）。舉例而言，ALU 252 可為經組態以執行算術運算、逐位元移位運算以及類似者的 16 位元 ALU、32 位元 ALU 或 64 位元 ALU。在各種實施例中，ALU 252 可包含電路系統以進行整數運算、浮點運算或兩者。舉例而言，ALU 252 可經組態以執行算術運算（諸如加法（+）、減法（-）、乘法（*）以及除法（÷））、逐位元運算（諸如及（&）、或（|）、異或（^）以及非（~）運算），以及張量運算。另外，在一些實施例中，ALU 252 可實施用於並行地對資料的向量進行運算的單指令多資料（single instruction multiple data；SIMD）或向量指令。由根據本揭露內容的實施例的 ALU 252 實施的向量運算的實例包含點積（ $\cdot$ ）、外積（ $\otimes$ ）、整流線性單元（rectified linear unit；ReLU）、平方（vsSqr）以及平方根（vsSqrt）。ALU 252 可用於基元運算及非基元運算。下表 1 列出由根據本揭露內容的一些實施例的 ALU 252 支援的運算。

表 1

ID	運算	描述
0	Rop = GIO	行讀取儲存於 Rop 中的資料
1	Rop = WIO	行寫入儲存於 Rop 中的資料
2	Rop = Rz	將 ALU 輸出 Rz 複製至 Rop
3	GIO = Rz	寫回至庫中
4	RIO = GIO	正常讀取至 DQ 輸出
5	RIO = Rz	將 Rz 驅動至 DQ 輸出
6	Rz = Rop (op) GIO	Rop 與來自庫的資料的運算
7	Rz = Rop (op) WIO	Rop 與廣播資料的運算
8	Rz = WIO (op) GIO	廣播資料與庫的運算
9	GIO = WIO	自 DQ 輸入正常寫入

【0052】 圖 2C 為根據本揭露內容的一個實施例的 DRAM 庫 200 的示意圖，所述 DRAM 庫 200 包含諸如上文所論述的彼等 DRAM 胞 210 陣列、列解碼器 220、IOSA 230、IMC 模組 250 以及行解碼器 240。

【0053】 圖 3 為根據本揭露內容的一個實施例的 DRAM 塊陣列的示意圖。在圖 3 中所繪示的實施例中，十六個 DRAM 庫 200 以四乘四陣列配置，且標記為 DRAM 庫 A 至 DRAM 庫 P 並配置成四個庫群組，所述庫群組標記為 BG0（包含 DRAM 庫 A、DRAM 庫 B、DRAM 庫 C 以及 DRAM 庫 D）、BG1（包含 DRAM 庫 E、DRAM 庫 F、DRAM 庫 G 以及 DRAM 庫 H）、BG2（包含 DRAM 庫 I、DRAM 庫 J、DRAM 庫 K 以及 DRAM 庫 L）以及 BG3（包含 DRAM 庫 M、DRAM 庫 N、DRAM 庫 O 以及 DRAM 庫 P）。如上文相對於圖 2C 所描述，圖 3 中所繪示的每一 DRAM 庫 200 可包含用於（例如，在不遍歷外部匯流排的情況下）在 DRAM 晶粒 120 內進行計算的 IMC 模組 250。此外，如圖 3 中所繪示，DRAM 晶粒 120 更包含經組態以將資料自外部源（例如，藉助於使 DRAM 晶粒 120 連接至多個其他堆疊的 DRAM 晶粒 120 及記憶體控制器 140 的矽穿孔或 TSV）多工至四個塊行的多工器 300（例如，4:1 多工器）。

舉例而言，多工器 300 可將 256 位元（256b）資料向量廣播至每一 DRAM 庫 200 或可用於將資料向量供應至 DRAM 庫 200 的特定行（行 BG0、行 BG1、行 BG2 或行 BG3）。

【0054】 包含整合至 DRAM 晶粒 120 中的記憶體內計算（IMC）（諸如 IMC 模組 250）的記憶體系統 100 加速記憶體有界主機運算的效能，此是由於資料不需要穿過外部匯流排（例如，匯流排 190）的瓶頸以進行計算。然而，IMC 仍可能遇到呈 ALU 管線及 DRAM 過程形式的計算負擔，以及呈資料置放及 DRAM 時序形式的記憶體負擔。

【0055】 因此，本揭露內容的實施例的態樣是關於用於將資料置放於 DRAM 內，以避免或減少 DRAM 時序負擔在由積體 IMC 模組 250 進行記憶體內計算時的影響的系統及方法。本揭露內容的實施例的一些態樣是關於用於達成經改良效能的軟體及硬體協同設計。

【0056】 在各種實施例中，記憶體控制器 140 提供運算且管理資料至 DRAM 晶粒 120 的輸入及資料自 DRAM 晶粒 120 的輸出。因此，本揭露內容的實施例的一些態樣是關於經組態以根據由主機處理器 170 提供至記憶體系統 100 的記憶體控制器 140 的指令將資料置放於 DRAM 晶粒 120 內的記憶體控制器 140。舉例而言，本揭露內容的實施例的一些態樣是關於提供用於與具有記憶體內計算及編譯器（例如，資料編譯器）的 HBM 交互的應用程式設計介面（application programming interface；API），所述編譯器經組態以生成命令以控制記憶體控制器 140 在使用 API 編譯或解譯程式的原始碼時根據對 API 作出的調用來置放資料。舉例而言，API

可提供用於進行通用矩陣-矩陣乘法（GEMM）的函數調用，且編譯器可基於包含待對資料進行的運算（例如，內積、外積、矩陣乘法以及類似者）、資料的大小（例如，資料是否在記憶體的字節內適配）的因素來生成命令序列，以控制記憶體控制器 140 以改良相對於資料的初級置放的效能的方式將表示運算元矩陣的資料置放至 DRAM 晶粒 120 中。當編寫軟體以經由 API 利用具有 IMC 的 DRAM 時，根據本揭露內容的一些實施例的編譯器或資料編譯器可將軟體的原始碼中的至少一些變換成由記憶體控制器 140 執行的命令，以將資料置放於 DRAM 晶粒 120 的特定方位中且控制 DRAM 晶粒 120 的 IMC 模組進行計算並儲存結果。

【0057】 為了論述，根據一個實施例，16 通路（lane）ALU 可達成半精確度浮點（floating-point；FP-16）計算的每秒 8 十億浮點運算（giga floating point operations per second；GFLOPS）的峰值效能。（儘管根據本揭露內容的實施例的 IMC 模組的效能在本文中相對於 FLOPS 進行描述，但本揭露內容的實施例不限於進行浮點運算，且當進行例如整數運算時各種資料佈局的相對效能可為類似的。）因此，使用四個晶粒（4H 或 4-Hi）（使用第二代高頻寬記憶體標準（second generation high bandwidth memory standard；HBM2））的堆疊的根據本發明的一個實施例的記憶體內計算（IMC）的實施方案可達成 FP-16 計算的每秒 8 兆浮點運算（tera floating point operations per second；TFLOPS）（每晶粒 256 個庫乘以 4 個晶粒的堆疊=1,024 個庫，每一庫具有對應 16 通路 ALU）。

【0058】 如以 TFLOPS 量測的峰值計算效能可在不同資料佈局情形下改變。在一個（第一）情況（標示 2OP）下，將來自記憶體模

組 110 外部的兩個運算元完全饋送至記憶體內計算 ALU，且完全緩衝及累加結果，由此產生上文論述的 8 TFLOPS 峰值計算效能。

【0059】 在另一（第二）情況（標示 1OP）下，將一個運算元自 HMB 外部完全饋送至 IMC，同時自 DRAM 晶粒 120 中的隨機方位讀取第二運算元。此導致峰值計算效能降低約 6.5 TFLOPS。

【0060】 在第三情況（標示 DR）下，兩個運算元位於 DRAM 的不同頁中且將結果寫回至 DRAM。在此情形下，所量測的計算效能約為 0.8 TFLOPS（例如，慢於 2OP 情況的數量級）。

【0061】 在第四情況（標示 SR）下，兩個運算元位於 DRAM 塊的相同列或相同頁中。此引起效能顯著改良至約 3.3 TFLOPS，優於 DR 情況。

【0062】 因此，可基於資料的考慮來作出各種資料佈局權衡。舉例而言，當在 1OP 情況下第一運算元位於 DRAM 中且第二運算元自 HBM 外部廣播時，效能較高（約 6.5 TFLOPS，如上文所提及），但此在主機處理器 170 上強加必須將第二運算元供應至 HBM 的負擔。

【0063】 圖 4A 示意性地示出根據本揭露內容的一個實施例的在一個運算元（矩陣 A）儲存於 DRAM 中且第二運算元（矩陣 B）自具有積體記憶體內計算（IMC）的記憶體模組的外部廣播的情況下用於通用矩陣-矩陣乘法（GEMM）的資料的置放。為了圖示，更詳細地繪示 DRAM 庫 200-O。如圖 4A 中所繪示，與第一運算元矩陣 A 相關聯的資料置放於 DRAM 庫 200-O 的一個頁 401（例如，第一列或第一頁）中，且與第二運算元矩陣 B 相關聯的資料（例如，經由多工器 300）自 DRAM 晶粒 120 的外部廣播。計算的

結果 C 可置放於 DRAM 庫 200-O 的不同頁 402 中。

【0064】 當兩個運算元位於 DRAM 中時，將運算元置放於相同頁或相同列（SR）中部分地藉由減少必須進行的 PRE 及 ACT 操作的數目來改良計算效能（例如，至約 3.3 TFLOPS），如在下文更詳細地論述，但強加關於將資料置放於 DRAM 的正確部分中的更多約束。

【0065】 圖 4B 示意性地示出根據本揭露內容的一個實施例的在兩個運算元（矩陣 A 及矩陣 B）儲存於具有積體 IMC 的記憶體模組的 DRAM 的相同頁中的情況下用於通用矩陣-矩陣乘法（GEMM）的資料的置放。為了圖示，更詳細地繪示 DRAM 庫 200-O。如圖 4B 中所繪示，與第一運算元矩陣 A 及第二運算元矩陣 B 兩者相關聯的資料均置放於 DRAM 庫 200-O 的頁 411（例如，第一列或第一頁）中。更詳細而言，頁 411 的第一半填充有來自第一運算元矩陣 A 的資料，且頁 411 的第二半填充有來自第二運算元矩陣 B 的資料。假定矩陣 A 及矩陣 B 與上文相對於圖 4A 所描述的彼等矩陣大小相同，則可能需要額外頁以儲存資料的其餘部分。因此，與第一運算元矩陣 A 及第二運算元矩陣 B 兩者相關聯的資料亦置放於頁 412 中。可接著將矩陣乘法的結果 C 儲存於 DRAM 庫 200-O 的頁 413 中。

【0066】 另一方面，將運算元置放於不同頁中更為靈活且減少對佈局的約束（例如，可能適合於具有未整齊地適配至記憶體的固定大小頁中的大小的資料），但通常可能以降低計算效能為代價。

【0067】 圖 4C 示意性地示出根據本揭露內容的一個實施例的在兩個運算元（矩陣 A 及矩陣 B）儲存於具有積體 IMC 的記憶體模

組的 DRAM 的不同頁中的情況下用於通用矩陣-矩陣乘法 (GEMM) 的資料的置放。為了圖示，更詳細地繪示 DRAM 庫 200-O。如圖 4C 中所繪示，與第一運算元矩陣 A 相關聯的資料置放於頁 421 中，與第二運算元矩陣 B 相關聯的資料置放於頁 422 中，且結果 C 置放於頁 423 中。

【0068】 上文相對於圖 4A、圖 4B 以及圖 4C 所論述的不同資料置放策略 (1OP、SR 以及 DR) 的各種效能影響將在下文相對於圖 5A、圖 5B、圖 6A、圖 6B、圖 7A 以及圖 7B 更詳細地描述。如例如圖 5A 中所繪示，矩陣 A 為 M 乘 K 矩陣且矩陣 B 為 K 乘 N 矩陣，因此矩陣 A 與矩陣 B 的乘積 (矩陣 C) 具有 M 乘 N 的尺寸。為了論述，以下實例描繪 $K = 5$ 的情況，但本揭露內容的實施例不限於此。根據標準矩陣乘法，藉由矩陣 A 的第一列 (位置 A_{ij} 中的每一者表示例如按水平次序的十六個半精確度浮點值的向量或「磚片 (tile)」，且圖 5A、圖 5B、圖 6A、圖 6B、圖 7A 以及圖 7B 描繪 A00、A01、A02、A03 以及 A04) 乘以矩陣 B 的第一行 (位置 B_{ij} 中的每一者表示例如按豎直次序的十六個半精確度浮點值的向量或「磚片」) 的成對乘法來計算出結果矩陣 C 的左上值 C00。換言之，C00 儲存 $A00 \cdot B00 + A01 \cdot B10 + A02 \cdot B20 + A03 \cdot B30 + A04 \cdot B40$ 。在圖 5A、圖 5B、圖 6A、圖 6B、圖 7A 以及圖 7B 中，使用陰影來識別共同相乘的運算元。更詳細而言，使用相同模式加陰影的兩個運算元共同相乘作為所示出計算的部分。儘管本揭露內容的實施例的態樣描述於經組態以對浮點運算元進行浮點運算的 IMC 的上下文中，但本揭露內容的實施例不限於此，且亦可應用於經組態以對整數運算元進行例如整數運算的 IMC。

【0069】圖 5A 為根據本揭露內容的一個實施例的一個運算元 (1OP) 資料佈局中的矩陣 A 的第一列乘以矩陣 B 的第一行的乘法的示意性描繪，其中一個運算元自外部饋送且一個運算元儲存於具有記憶體內計算的 DRAM 庫中。如圖 5A 中所繪示，矩陣 A 的第一列的磚片 A00、磚片 A01、磚片 A02、磚片 A03 以及磚片 A04 儲存於 DRAM 庫 200 的相同列 401 中，矩陣 B 的向量或磚片 B00、磚片 B10、磚片 B20、磚片 B30 以及磚片 B40 自外部供應，且結果 (例如，C00) 儲存於 DRAM 庫 200 的另一頁 402 中。

【0070】計算乘積的過程可藉由自 DRAM 庫 200 讀取值 A00 來計算 $A00 \cdot B00$ 開始。此將涉及打開頁 401，且因此將需要預充電 (PRE) 命令以準備感測放大器 232，繼之以激活 (ACT) 以將頁 401 載入至感測放大器 232 中，以及 READ 命令以自感測放大器 232 將 A00 的值載入至 IMC 模組 250 中。如上文所提及，B00 來自外部作為輸入，且因此不需要用以檢索此值的 DRAM 運算。ALU 252 可接著計算乘積 $A00 \cdot B00$ 且將臨時結果儲存於輸出緩衝器 (例如，輸出暫存器 Rz) 中。

【0071】接下來，ALU 252 藉由自 DRAM 讀取 A01 來計算乘積 $A01 \cdot B10$ 。此亦需要 PRE 命令、ACT 命令以及 READ 命令。向量或磚片 B10 自外部提供作為輸入，且因此 $A01 \cdot B10$ 可接著計算出且加至儲存 $A00 \cdot B00$ 的緩衝器 (例如，輸出暫存器 Rz) 中的臨時結果。接著可針對矩陣 A 及矩陣 B 的剩餘值重複此過程。因此，每一計算 (例如，諸如磚片 A00 及磚片 B00 的兩個運算元的乘法) 需要每計算一個 PRE、一個 ACT 以及一個 READ。如在下文相對於圖 8 更詳細地論述，在一些實施例中，IMC 更包含經組態以儲

存值且用先前經儲存值與新接收到的值的總和更新經儲存值的累加器。

【0072】圖 5B 為根據本揭露內容的一個實施例的 1OP 資料佈局中的矩陣 A 的第一列的第一值乘以矩陣 B 的每一行的第一值的乘法（其中資料再用）的示意性描繪，其中一個運算元自外部饋送且一個運算元儲存於具有記憶體內計算的 DRAM 庫中。圖 5B 中所繪示的過程不同於圖 5A 中所繪示的過程，不同之處在於針對矩陣 B 的不同行再用自 DRAM 載入的（例如，儲存於運算元暫存器 Rop 254 中的）資料。特定而言，當兩個矩陣相乘時，矩陣 A 的每一列的第 j 元素與矩陣 B 的第 j 列的每一元素相乘。因此，藉由載入矩陣 A 的每一元素一次且將其乘以矩陣的所有 N 個行，自 DRAM 庫 200 載入資料的成本跨 N 個行分攤。更詳細而言，當矩陣 A 乘以矩陣 B 時，如前所述，可使用 PRE 命令、ACT 命令以及 READ 命令自 DRAM 庫 200 讀取磚片 A00，且可自外部接收到磚片 B00 作為輸入。ALU 252 計算 $A00 \cdot B00$ 以計算 C00 的總和的一部分。然而，可再次使用磚片 A00 且乘以 B01（自外部接收到）以計算 $A00 \cdot B01$ （其為待針對 C01 計算出的乘積中的一者），而非自 DRAM 載入磚片 A01（例如，使用另一 PRE、ACT 以及 READ 序列）。因此，每一計算需要每計算 $1/N$ 個 PRE、 $1/N$ 個 ACT 以及一個 READ（此是由於用於載入矩陣 A 的另一部分的 PRE 命令及 ACT 命令可在矩陣 B 的 N 個行上分攤）。

【0073】圖 6A 為根據本揭露內容的一個實施例的相同列（SR）資料佈局中的矩陣 A 的第一列乘以矩陣 B 的第一行的乘法的示意性描繪，其中兩個運算元儲存於具有記憶體內計算的 DRAM 庫的相

同頁中。如圖 6A 中所繪示，矩陣 A 的第一列的磚片 A00、磚片 A01 以及磚片 A02 及矩陣 B 的第一行的磚片 B00、磚片 B10 以及磚片 B20 儲存於 DRAM 庫 200 的相同頁 411 中，且矩陣 A 的磚片 A03 及磚片 A04 以及矩陣 B 的磚片 B30 及磚片 B40 儲存於 DRAM 庫 200 的頁 412 中。結果（例如，C00）儲存於 DRAM 庫 200 的頁 413 中。

【0074】計算結果（例如，內積）的過程可藉由自 DRAM 庫 200 讀取磚片 A00 來計算 $A00 \cdot B00$ 開始。此將涉及打開頁 411，且因此將需要預充電(PRE)命令，繼之以激活(ACT)命令，以及 READ 命令以自感測放大器 232 將磚片 A00 載入至 IMC 模組 250 中。可接著自 DRAM 讀取磚片 B00。然而，由於磚片 B00 與 A00 位於相同頁 411 中，故所述磚片 B00 的值已儲存於感測放大器 232 中，且因此，READ 命令足夠（不需要進一步進行 PRE 及 ACT 以將磚片 B00 讀取至 IMC 模組 250 中）。因此，一旦讀取磚片 A00 及磚片 B00，則 ALU 252 計算 $A00 \cdot B00$ ，且可將臨時結果儲存於緩衝器中。類似地，同樣可在無 PRE 及 ACT 的情況下使用 READ 命令來進行讀取磚片 A01 及磚片 B10 以計算 $A01 \cdot B10$ ，此是由於磚片 A01 及磚片 B10 均亦位於頁 411 中，且因此在最初打開頁 411 以讀取 A00 時先前儲存於感測放大器 232 中。因此，每一計算採用每計算 $1/r$ 個 PRE、 $1/r$ 個 ACT 以及 2 個 READ 操作，其中 r 為儲存於 DRAM 庫 200 的相同頁中的值的匹配對的數目。舉例而言，如上文所提及，圖 6A 描繪矩陣 A 的第一列的磚片 A00、磚片 A01 以及磚片 A02 以及矩陣 B 的第一行的磚片 B00、磚片 B10 以及磚片 B20 儲存於 DRAM 庫的頁 411 中的情況。因此，計算 $A00 \cdot B00$ 、

計算 $A01 \cdot B10$ 以及計算 $A02 \cdot B20$ 將各自採用（分攤） $1/3$ 個 PRE 命令、 $1/3$ 個 ACT 命令以及 2 個 READ 命令，此是由於頁 411 含有三對磚片（例如， $r = 3$ ）。當計算 $A03 \cdot B30$ 及 $A04 \cdot B40$ 時，此等計算將各自採用（分攤） $1/3$ 個 PRE 命令、 $1/3$ 個 ACT 命令以及 2 個 READ 命令，此是由於頁 412 含有兩對值（例如， $r = 2$ ）。每計算所需的 PRE 命令及 ACT 命令的數目的此減少改良整個計算的效能。

【0075】圖 6B 為根據本揭露內容的一個實施例的 SR 資料佈局中的矩陣 A 的第一列的第一值乘以矩陣 B 的每一行的第一值的乘法（其中資料再用）的示意性描繪，其中兩個運算元儲存於具有記憶體內計算的 DRAM 庫的相同頁中。以類似於圖 5A 及圖 5B 的配置之間的比較的方式，圖 6B 中所繪示的計算過程不同於圖 6A 中所繪示的計算過程，不同之處在於再用自 DRAM 載入的值。更詳細而言，基於矩陣乘法涉及第一運算元的給定列的第 i 元素與第二運算元的每一行的第 i 元素相乘的理解，矩陣 A 及矩陣 B 的元素配置於 DRAM 庫 200 中，使得矩陣 A 的給定列的每一元素的資料與其將乘以的矩陣 B 的列的值置放於相同頁中。

【0076】舉例而言，如圖 6B 中所繪示，當矩陣 A 乘以矩陣 B 時，矩陣 A 的磚片 A00 將在計算結果矩陣 C 的第一列的部分的過程中乘以矩陣 B 的每一行的第一磚片（矩陣 B 的第一列的每一元素，例如，B00、B01、B02、B03、B04、B05.....如圖 6B 中所繪示）（例如，作為 C00、C01、C02、C03、C04、C05.....的項中的一者）。同樣，矩陣 A 的值 A01 將乘以矩陣 B 的每一行的第二磚片（矩陣 B 的第二列的每一元素，例如，B10、B11、B12、B13、B14、B15.....

如圖 6B 中所繪示)。

【0077】 因此，矩陣 A 的至少一個值與矩陣 B 的對應值儲存於相同頁中。在圖 6B 中所繪示的具體實例中，頁 411 儲存來自矩陣 A 的磚片 A00 及來自矩陣 B 的磚片 B00、磚片 B01、磚片 B02、磚片 B03、磚片 B04、磚片 B05.....，且頁 412 儲存來自矩陣 A 的磚片 A01 及來自矩陣 B 的磚片 B10、磚片 B11、磚片 B12、磚片 B13、磚片 B14、磚片 B15.....。歸因於資料在 DRAM 內的此配置，假定 DRAM 的頁可儲存至少 $N+1$ 個條目，則每一計算將採用 $1/N$ 個 PRE 命令、 $1/N$ 個 ACT 命令以及 $(N+1)/N$ 個 READ 命令（此是由於 PRE 命令及 ACT 命令可在矩陣 B 的 N 個行上分攤）。每計算所需的 PRE 命令及 ACT 命令的數目的此減少改良整個計算的效能。

【0078】 圖 7A 為根據本揭露內容的一個實施例的不同列（DR）資料佈局中的矩陣 A 的第一列乘以矩陣 B 的第一行的乘法的示意性描繪，其中運算元儲存於具有記憶體內計算的 DRAM 庫的不同頁中。如圖 7A 中所繪示，矩陣 A 的第一列的磚片 A00、磚片 A01、磚片 A02、磚片 A03 以及磚片 A04 儲存於頁 421 中，而矩陣 B 的第一行的磚片 B00、磚片 B10、磚片 B20、磚片 B30 以及磚片 B40 儲存於不同頁 422 中。

【0079】 計算 C00 的過程可開始於藉由使用命令的 PRE、ACT 以及 READ 序列自 DRAM 庫 200 的列 421 讀取 A00，繼之以使用命令的 PRE、ACT 以及 READ 序列讀取磚片 B00 來計算 $A00 \cdot B00$ ，此是由於磚片 B00 與磚片 A00 位於不同列 422 中。ALU 252 可接著計算 $A00 \cdot B00$ 且將結果儲存於臨時緩衝器中。為繼續計算 C00，

可自列 421 讀取磚片 A01，繼之以自列 422 讀取磚片 B10，其中每一值涉及對 DRAM 庫 200 進行 PRE、ACT 以及 READ 序列。因此，每一計算採用 2 個 PRE 命令、2 個 ACT 命令以及 2 個 READ 命令。

【0080】圖 7B 為根據本揭露內容的一個實施例的 DR 資料佈局中的矩陣 A 的第一列的第一值乘以矩陣 B 的每一行的第一值及矩陣 A 的第一列的第二值乘以矩陣 B 的每一行的第二值的乘法（其中資料再用）的示意性描繪，其中運算元儲存於具有記憶體內計算的 DRAM 庫的不同頁中。

【0081】以類似於上文相對於圖 5B 所描述且在 1OP 中資料再用的方式，在進行與矩陣 B 的乘法中的每一者時再用自矩陣 A 檢索到的資料可使得記憶體操作的數目減少。如上文相對於圖 7A 所論述，矩陣 A 的給定列的每一第 i 值乘以矩陣 B 的第 i 列中的每一值。因此，若將矩陣 B 的給定列中的所有值儲存於 DRAM 庫 200 的相同頁中，則可減少 PRE 命令及 ACT 命令的數目，由此改良效能。

【0082】舉例而言，將矩陣 A 乘以矩陣 B 的過程可開始於將矩陣 A 的磚片 A00 乘以矩陣 B 的第一列中的 N 個磚片（磚片 B00、磚片 B01、磚片 B02.....磚片 B0N）中的每一者以計算 N 個部分和（例如，C00、C01、C02.....C0N 的部分）。此過程可藉由自 DRAM 庫 200 載入磚片 A00 開始，此可涉及 PRE 命令、ACT 命令以及 READ 命令以打開列 421，且將矩陣 A 的磚片 A00 載入至運算元暫存器 Rop 254 中。在載入 A00 之後，可載入矩陣 B 的第一列的磚片 B00、磚片 B01、磚片 B02.....磚片 B0N 以乘以磚片 A00。如

圖 7B 中所繪示，當矩陣 B 的此等值均位於相同頁中（例如，B00、B01、B02、B03、B04.....位於圖 7B 中所繪示的 DRAM 庫 200 的頁 422 中）時，則用以存取此頁的 PRE 命令及 ACT 命令可跨儲存於頁中的 N 個值分攤。因此，將來自矩陣 A 的一個值乘以矩陣 B 的一個列中的每一值需要 2 個 PRE 命令、2 個 ACT 命令以及 N+1 個 READ 命令，此在跨列的 N 個值分攤時產生每計算 2/N 個 PRE 命令、2/N 個 ACT 命令以及 (N+1)/N 個 READ 命令。以類似於上文給出的實例的方式，DRAM 命令的數目的此減少改良整個計算（例如，矩陣乘法運算）的效能。

【0083】 使用循環級高頻寬記憶體-記憶體內函數 (high bandwidth memory-function in memory; HBM-FIM) 模擬器以實驗方式測試上述資料置放選項，所述 HBM-FIM 模擬器是針對具有通用矩陣-矩陣乘法 (GEMM) 跡線的記憶體內計算 (IMC) 定製的。實驗架構包含具有四個 4H HBM2 模組的 HBM，其中 DRAM 具有 16,384 (16Kib) 個列乘以 8,192 (8Kib) 個行的陣列大小（例如，每一頁的大小為 8,192b），且列緩衝器具有 8,192 個位元 (8Kib) 的大小。IMC 模組包含具有潛時循環的 16 通路 FP-16 向量單元及 768 位元緩衝器，且能夠進行管線式操作。在各種實施例中，緩衝器可更小（例如，512 個位元，包含 256 位元輸入緩衝器及 256 位元輸出緩衝器）或可更大（例如，具有 256 位元輸入緩衝器及 1,024 位元輸出緩衝器的 1,280 位元緩衝器）。在一些實施例中，輸入緩衝器大於 256 個位元（例如，512 個位元）。

【0084】 本揭露內容的實施例的一些態樣是關於在 DRAM 庫 200 的層級上包含額外緩衝器及累加器（例如，在 DRAM 庫處在每庫

基礎上設置的額外硬體)。

【0085】 圖 8 為根據本揭露內容的一個實施例的 DRAM 庫的記憶體內計算 (IMC) 模組的示意性方塊圖，其中 IMC 模組進一步連接至結果緩衝器、累加器以及緩衝器。

【0086】 如圖 8 中所繪示，ALU 252 可接收輸入運算元 A (在圖 8 中標記為 256 位元運算元 A[0:255]) 及輸入運算元 B (在圖 8 中標記為 256 位元運算元 B[0:255])。ALU 對兩個輸入運算元進行運算 (例如，相加、相乘、點積、外積等) 且計算結果 C (在圖 8 中標記為 256 位元結果 C[0:255])。

【0087】 如圖 8 中所繪示，IMC 模組 250 可更包含連接至 ALU 252 的輸出的累加器 802。舉例而言，ALU 252 的輸出可連接至結果暫存器 Rz 256 及累加器 802 兩者，使得結果 C 儲存於結果暫存器 Rz 256 中且供應至累加器 802。累加器 802 包含儲存經累加值 (例如，256 位元值) 的累加器暫存器。當累加器 802 自 ALU 252 接收到新的結果時，將新的結果加至已儲存於累加器暫存器中的經累加值 (例如，儲存於累加器 802 的累加器暫存器中的值經更新或設定為新的結果與先前儲存於累加器 802 中的值的總和)。在一些實施例中，累加器 802 可經組態以回應於重設命令而重設累加器暫存器 (例如，以將儲存於累加器暫存器中的經累加值設定為零)。當計算內積或點積 (諸如當計算矩陣乘法時) 時 (例如，在結果矩陣中的每一值為第一運算元的一列與第二運算元的一行的內積的情況下)，累加器 802 可尤其有幫助。在圖 8 中所繪示的實施例中，IMC 模組 250 更包含 1,024 位元輸出緩衝器，繪示為第一 512 位元緩衝器 812 及第二 512 位元緩衝器 814。第一 512 位元緩衝器

812 及第二 512 位元緩衝器 814 可以接收累加器 802 的輸出（例如， $D[0:16]$ ）。能夠儲存多個結果值的更大輸出緩衝器允許記憶體控制器 140 控制 DRAM 庫以在打開 DRAM 陣列的另一頁以儲存多個結果之前立即計算彼等結果。舉例而言，在如圖 6B 中所繪示的 SR 資料佈局中的資料再用的情況下，來自第一運算元矩陣 A 的一個列的值乘以來自第二運算元矩陣 B 的一列的不同對應值，以計算寫入至 DRAM 庫 200 的與儲存運算元的頁（例如，頁 411）不同的頁（例如，如圖 6A 及圖 6B 中所繪示的頁 413）的部分和，此可能需要額外 PRE 命令及 ACT 命令以打開用於儲存結果的頁。然而，更大輸出緩衝器減少完成運算所需的頁切換量（由於每一頁切換需要 PRE 及 ACT），由此改良計算的效能。儘管圖 8 描繪具有累加器 802 以及更大輸出緩衝器 812 及更大輸出緩衝器 814 兩者的 IMC 模組 250，但本揭露內容的實施例不限於此，且亦可包含其中 IMC 模組 250 包含累加器 802 而不包含額外輸出緩衝器 812 及輸出緩衝器 814 的實施例，以及其中 IMC 模組 250 包含額外輸出緩衝器 812 及輸出緩衝器 814 而不包含累加器 802 的實施例。在本揭露內容的一些實施例中，IMC 模組 250 可包含平行配置的多個累加器 802（例如，使得可同時累加多個值）。本揭露內容的實施例進一步不限於具有兩個 512 位元輸出緩衝器 812 及輸出緩衝器 814 的情況，而是可包含大於或小於 512 個位元的緩衝器及/或大於兩個輸出緩衝器或少於兩個輸出緩衝器。IMC 模組 250 的各種部分可具有不同潛時量。舉例而言，當進行乘法運算的 ALU 252 可引入四個循環潛時時，累加運算可涉及一個循環潛時。

【0088】 本揭露內容的實施例的態樣將在下文在實施通用矩陣-矩

陣乘法（GEMM）的上下文中更詳細地描述。

【0089】圖 9 為根據本揭露內容的一些實施例的用以解釋使用相同列（SR）佈局的計算的通用矩陣-矩陣乘法（GEMM）的實例的示意性圖示。更詳細而言，圖 9 描繪矩陣 A 乘以矩陣 B 的乘法以及藉由將矩陣 A 與矩陣 B 的乘積加至當前儲存於累加器 802 中的值來更新儲存於累加器（例如，累加器 802）中的結果 C（結果 $C += \text{矩陣 A} \times \text{矩陣 B}$ ）。在圖 9 中所繪示的計算中，矩陣 A 的 16 磚片與矩陣 B 的 16 磚片共同相乘，其中每一磚片具有 16 個 FP-16 元素（256 位元）。特定內部磚片組織佈局將在下文更詳細地論述，但不同佈局的 GEMM 的整體複雜度保持相同。如圖 9 中所標示，以類似圖案加陰影的磚片為共同相乘的磚片。如上文所提及，圖 9 描繪相同列（SR）資料佈局，其中矩陣 A 及矩陣 B 的所有三十二個值儲存於 DRAM 庫 200 的相同頁 414 中，使得每一計算的兩個運算元（例如，磚片 A00 以及磚片 B00、磚片 B01、磚片 B02 以及磚片 B03）均位於相同頁 414 中。個別計算的結果可在結果暫存器 Roz 256 中累加以計算結果 C。

【0090】作為磚片級乘法的一個具體實例，當計算自 DRAM 庫 200 的相同頁載入的兩個磚片 A00 與磚片 B00 的內積（例如，點積）時，磚片 A00 可為矩陣 B 的加點磚片 B00。如上文所論述，在一些實施例中，每一磚片包含十六個 FP-16 元素。舉例而言，磚片 A00 可包含元素 a00、元素 a01.....元素 a15，且磚片 B00 可包含元素 b00、元素 b01.....元素 b15，且磚片 A00 與磚片 B00 的內積為兩兩乘積的總和： $a00*b00 + a01*b01 + + a15*b15$ 。因此，兩個磚片的內積（或點積）產生單一元素或單一值，且因此根據本揭

露內容的實施例的記憶體內計算 (IMC) 模組可用用於儲存結果的更少緩衝器暫存器 (例如, 累加器暫存器 802) 計算內積。因此, 包含累加器暫存器的根據本揭露內容的一些實施例的 IMC 模組可良好地適合於進行涉及值的累加的計算, 諸如計算內積。在一些實施例中, 適合於計算內積的具有累加器的 IMC 模組用於使用經訓練神經網路來進行推理 (或轉送傳播), 諸如其中第一運算元可表示輸入 (例如, 來自前一層的激活) 且第二運算元可表示與經訓練神經網路的一層的神經元相關聯的權重。

【0091】 作為磚片級乘法的另一實例, 當計算自 DRAM 庫 200 的相同頁載入的兩個磚片的外積時, 將第一磚片的每一值乘以第二磚片的每一值, 且可同時載入兩個磚片。舉例而言, 在每一磚片包含十六個值的情況下, 如上文所描述, 兩個磚片的外積具有 $16 \times 16 = 256$ 個輸出值。外積可比內積用更簡單的硬體來計算, 此是由於不需要樹加法器累加, 且是由於可並行計算 256 個輸出值。然而, 計算外積需要根據本揭露內容的實施例的記憶體內計算模組 250 中的許多輸出緩衝器暫存器 (例如, 輸出緩衝器 812 及輸出緩衝器 814) (例如, 以便儲存外積的結果的所有值)。舉例而言, 假定每一磚片包含最大十六個 FP-16 值, 則經組態以計算外積的 IMC 模組 250 可包含足夠大以儲存 256 個值的輸出緩衝器 (例如, 256×16 位元 = 4,096 位元緩衝器)。因此, 包含具有用於儲存大於或等於第一運算元及第二運算元的磚片中的值的數目的乘積的輸出緩衝器 (在以上實例中, 輸出緩衝器儲存至少 256 個值, 此是由於運算元的磚片中的每一者包含 16 個值) 的根據本揭露內容的一些實施例的 IMC 模組可良好地適合於進行涉及運算元的外積的計

算。

【0092】 作為磚片級乘法的第三實例，當計算自 DRAM 庫 200 的相同頁載入的兩個磚片的張量積時，兩個磚片可包含以矩陣配置的值。舉例而言，在每一磚片包含十六個 FP-16 值的情況下，如上文所描述，每一磚片可配置成值的 4 乘以 4 矩陣。計算此等磚片的張量積生成值的 4 乘以 4 結果矩陣。舉例而言，假定每一磚片包含最大十六個 FP-16 值，則經組態以計算兩個 4 乘以 4 磚片的 16 個值的張量積的 IMC 模組 250 可具有足夠大以儲存 16 個值的輸出緩衝器（例如， $16 * 16$ 位元 = 256 位元緩衝器）。因此，包含足夠大以儲存等於第一運算元及第二運算元中的更大者中的值的數目的值的數目的輸出緩衝器的根據本揭露內容的實施例的 IMC 模組可良好地適合於進行涉及運算元的張量積的計算。相較於適合於計算內積及計算外積的 IMC 模組硬體，在根據本揭露內容的實施例的記憶體內計算模組中實施張量積呈現具有更簡單硬體及中間數目個緩衝器暫存器的中間基礎。

【0093】 更詳細而言，在兩個 4 乘以 4 矩陣 A 及 4 乘以 4 矩陣 B 之間進行張量積涉及對矩陣 A/矩陣 B 的 64 個乘法運算、48 個加法運算以及轉置。當使用根據本揭露內容的一個實施例的 16 通路 e-ALU 252 時，可並行進行 16 個運算。因此， $64 + 48 = 112$ 個運算可在 ALU 的七個循環中進行（假定乘法可在一個循環中進行）。在本揭露內容的另一實施例中，為提高效能，可使用 64 通路 ALU 252 來計算表示 2 至 3 個循環或 4 奈秒至 6 奈秒中的兩個 4 乘以 4 矩陣的兩個磚片的張量積。因此，增加 ALU 252 中的通路的數目可增加計算之間可能的並行化的量，由此亦提高效能。

【0094】 因此，本揭露內容的實施例的態樣是關於用於具有記憶體內計算的動態隨機存取記憶體（DRAM）系統中的資料置放的系統及方法。在 SR（單一頁）資料佈局的上下文中計算內積、外積以及張量積的以上三個實例中，單一系列包含與具有（矩陣的）行資料的第二磚片位於相同頁中的具有（矩陣的）列資料的第一磚片。

【0095】 根據本揭露內容的一些實施例，記憶體模組 110 的記憶體控制器 140（或客戶端側記憶體控制器）根據自主機處理器 170 接收到的指令來控制自主機處理器 170 接收到的資料（運算元）的置放。舉例而言，基於指令為計算點積、外積抑或張量積，且基於運算元的大小（例如，矩陣的尺寸），記憶體控制器 140 可使用 1OP、SR 或 DR 資料佈局且控制 DRAM 庫相應地儲存資料。

【0096】 圖 10 為根據本揭露內容的實施例的用於控制資料在具有記憶體內計算模組的 DRAM 庫中的置放的方法的流程圖。在操作 1110 中，記憶體控制器 140 接收運算元（例如，第一運算元及第二運算元）及待應用於運算元的指令（例如，函數）。此等指令可包含例如計算內積、外積、使用經訓練神經網路及輸入向量來進行推理、進行用於訓練神經網路的反向傳播演算法的步驟以及類似者。

【0097】 在操作 1130 中，記憶體控制器 140 基於指令來判定使用哪一資料佈局（例如，1OP、SR 或 DR）以將一個或兩個運算元儲存於 DRAM 晶粒上的 DRAM 庫中。在一些實施例中，（例如，來自主機記憶體控制器 180 的）指令明確指定使用哪一資料佈局。在一些實施例中，基於將進行的計算的類型以及用於加速 DRAM 庫中的此類計算的記憶體或硬體要求來選擇佈局（例如，計算點積

的指令可使得將資料置放於具有累加器的 DRAM 庫中，而計算外部積的指令可使得將資料置放於具有更多輸出緩衝器的 DRAM 庫中)。

【0098】 在操作 1150 中，記憶體控制器 140 基於所選佈局將第一運算元及第二運算元供應至 DRAM 庫。舉例而言，在 1OP 情況下，記憶體控制器 140 控制 DRAM 庫 200 至少儲存第一運算元的第一磚片，且將第二運算元的第二磚片直接供應至 DRAM 庫的 IMC 模組。作為另一實例，在 SR 情況下，記憶體控制器 140 控制 DRAM 庫 200 將對應於第一運算元及第二運算元的磚片儲存於 DRAM 庫的相同列或相同頁中。

【0099】 在操作 1170 中，記憶體控制器 140 控制 DRAM 庫 200 的記憶體內計算模組基於輸入指令來進行運算。舉例而言，在使用經訓練模型來進行推理的指令的情況下，運算可包含基於一個運算元來準備輸入向量以及基於第二運算元中的經儲存參數對輸入向量的值進行加權。

【0100】 儘管本揭露內容已呈現某些例示性實施例，但應理解，本揭露內容不限於所揭露實施例，而是相反，意欲覆蓋包含於隨附申請專利範圍及其等效物的精神及範圍內的各種修改及等效配置。

【符號說明】

【0101】

100:記憶體系統

110:記憶體模組

120:動態隨機存取記憶體晶粒

130:內部記憶體匯流排

140:記憶體控制器

170:主機處理器

180:主機記憶體控制器

190:外部匯流排

200、200-O、A、B、C、D、E、F、G、H、I、J、K、L、M、

N、O、P:DRAM 庫

210:DRAM 胞

212:電容器

214:開關

220:列解碼器

230:輸入/輸出感測放大器層

232:感測放大器

234、300:多工器

236:全域 IO 層

240:行解碼器

250:記憶體內計算模組

252:算術邏輯單元

254:運算元暫存器

256:結果暫存器

257:第一多工器

258:第二多工器

260:輸入-輸出模組

401、402、411、412、413、414、421、422、423:頁

802:累加器

812:第一 512 位元緩衝器

814:第二 512 位元緩衝器

1110、1130、1150、1170:操作

a00、a01、a15、b00、b01、b15、A00、A01、A02、A03、A04、
B00、B01、B02、B03、B04、B05、B0N、B10、B11、B12、B13、
B14、B15、B20、B30、B40:磚片

Aij、Bij:位置

A[0:255]、B[0:255]:256 位元運算元

B1、B8192、Bi、Bm:位元線

BG0、BG1、BG2、BG3: 庫群組

C00、C01、C02、C03、C04、C05、C0N:值

C[0:255]:256 位元結果

R1、Rj、Rn:列賦能線

【發明申請專利範圍】

【請求項1】 一種記憶體模組，包括：

記憶體晶粒，包括多個動態隨機存取記憶（DRAM）庫，所述多個 DRAM 庫中的每一者包括：

DRAM 胞陣列，以多個頁配置，所述多個頁中的每一者包括多個 DRAM 胞，所述多個 DRAM 胞中的每一者儲存位元值；

列緩衝器，經組態以儲存所述多個頁中的打開頁的值；

輸入/輸出（IO）模組；以及

記憶體內計算（IMC）模組，包括：

算術邏輯單元（ALU），經組態以自所述列緩衝器或所述 IO 模組接收運算元，以及基於所述運算元及多個算術邏輯運算中的所選算術邏輯運算來計算輸出；以及

結果暫存器，經組態以儲存由所述 ALU 計算出的所述輸出；以及

記憶體控制器，經組態以：

自主機處理器接收第一運算元、第二運算元以及指令；

基於所述指令自多個資料佈局判定資料佈局；

根據所述資料佈局將所述第一運算元及所述第二運算元供應至所述多個 DRAM 庫；以及

控制所述多個 DRAM 庫的所述 IMC 模組以根據所述指令對所述第一運算元及所述第二運算元進行所述多個算術邏輯運算中的算術邏輯運算。

【請求項2】 如請求項 1 所述的記憶體模組，其中所述多個資料佈局包括：

一個運算元 (1OP) 資料佈局，其中將所述第一運算元寫入至所述多個 DRAM 胞且將所述第二運算元自所述主機處理器直接供應至所述多個 DRAM 庫的所述 IMC 模組。

【請求項3】 如請求項 2 所述的記憶體模組，其中所述 IMC 模組更包括運算元暫存器，且

其中所述記憶體控制器進一步經組態以：

將所述第一運算元的第一磚片儲存於所述運算元暫存器中；
以及

對儲存於所述運算元暫存器中的所述第一運算元及所述第二運算元的多個第二磚片中的每一者進行所述算術邏輯運算。

【請求項4】 如請求項 1 所述的記憶體模組，其中將所述第一運算元劃分成多個第一磚片且將所述第二運算元劃分成多個第二磚片，所述多個第一磚片中的每一者以及所述多個第二磚片中的每一者包括多個值，且

其中所述多個資料佈局包括相同頁 (SP) 資料佈局，其中所述記憶體控制器將所述多個第一磚片中的一或多者及所述多個第二磚片中的一或多者儲存於所述多個 DRAM 胞的相同頁中。

【請求項5】 如請求項 4 所述的記憶體模組，其中所述 IMC 模組更包括運算元暫存器，且

其中所述記憶體控制器進一步經組態以：

將所述多個第一磚片中的一或多者中的第一磚片儲存於所述運算元暫存器中；以及

對儲存於所述運算元暫存器中的所述第一運算元及與所述第一磚片儲存於所述 DRAM 胞陣列的所述相同頁中的所述多個第二

磚片中之一或多者中的每一者進行所述算術邏輯運算。

【請求項6】 如請求項 5 所述的記憶體模組，其中所述多個 DRAM 庫的所述 IMC 模組更包括累加器，所述累加器包括經組態以儲存經累加值的累加器暫存器，所述累加器經組態以：

接收由所述 ALU 計算出的所述輸出；以及

用所述經累加值與所述輸出的總和更新所述累加器暫存器，

且

其中所述指令包括計算所述第一運算元與所述第二運算元的內積，

其中所述多個第一磚片中的所述第一磚片儲存列資料，

其中所述多個第二磚片中的第二磚片包括行資料。

【請求項7】 如請求項 5 所述的記憶體模組，其中所述第一磚片具有第一數目個值且所述多個第二磚片中的第二磚片具有第二數目個值，

其中所述多個 DRAM 庫的所述 IMC 模組包括輸出緩衝器，

其中所述輸出緩衝器具有用於儲存大於或等於所述第一數目個值與所述第二數目個值的乘積的大小，且

其中所述指令包括計算所述第一運算元與所述第二運算元的外積，

其中所述多個第一磚片中的所述第一磚片儲存列資料，

其中所述多個第二磚片中的所述第二磚片包括行資料。

【請求項8】 如請求項 5 所述的記憶體模組，其中所述第一磚片具有第一數目個值且所述多個第二磚片中的第二磚片具有第二數目個值，

其中所述多個 DRAM 庫的所述 IMC 模組包括輸出緩衝器，所述輸出緩衝器具有用於儲存大於或等於所述第一數目個值及所述第二數目個值中的更大者的大小，且

其中所述指令包括計算所述第一運算元與所述第二運算元的張量積，其中所述多個第一磚片中的所述第一磚片儲存列資料且所述第二磚片中的所述第二磚片包括行資料。

【請求項9】 如請求項 1 所述的記憶體模組，其中將所述第一運算元劃分成多個第一磚片，且

其中將所述第二運算元劃分成多個第二磚片，

其中所述多個第一磚片中的每一者以及所述多個第二磚片中的每一者包括多個值，且

其中所述多個資料佈局包括不同頁（DP）資料佈局

其中所述記憶體控制器儲存所述 DRAM 胞陣列的第一頁中的所述多個第一磚片的子集及所述 DRAM 胞陣列的第二頁中的所述多個第二磚片的子集。

【請求項10】 如請求項 9 所述的記憶體模組，其中所述 IMC 模組更包括運算元暫存器，且

其中所述記憶體控制器進一步經組態以：

將來自所述第一頁的所述第一運算元的第一磚片儲存於所述運算元暫存器中；以及

對儲存於所述運算元暫存器中的所述第一運算元及來自所述第二頁的所述第二運算元的所述多個第二磚片中的每一者進行所述算術邏輯運算。

【請求項11】 如請求項 1 所述的記憶體模組，所述多個 DRAM 庫

中的每一者的所述 IMC 模組更包括經組態以緩衝由所述 ALU 計算出的所述輸出的緩衝器。

【請求項12】如請求項 11 所述的記憶體模組，其中所述緩衝器為所述結果暫存器的大小的至少四倍。

【請求項13】如請求項 1 所述的記憶體模組，所述多個 DRAM 庫中的每一者的所述 IMC 模組更包括累加器，所述累加器包括經組態以儲存經累加值的累加器暫存器，所述累加器經組態以：

接收由所述 ALU 計算出的所述輸出；以及

用所述經累加值與所述輸出的總和更新所述累加器暫存器。

【請求項14】如請求項 1 所述的記憶體模組，其中所述記憶體模組為包括藉由矽穿孔連接的記憶體晶粒堆疊的高頻寬記憶體（HBM）模組，

其中所述記憶體晶粒堆疊包括所述記憶體晶粒。

【請求項15】一種用於在記憶體內進行計算的方法，包括：

由記憶體模組的記憶體控制器接收第一運算元、第二運算元、指令；

由所述記憶體控制器基於所述指令自多個資料佈局判定資料佈局；

根據所述資料佈局將所述第一運算元及所述第二運算元供應至所述記憶體模組的一或多個動態隨機存取記憶（DRAM）庫；以及

控制所述一或多個 DRAM 庫的記憶體內計算（IMC）模組以根據所述指令對所述第一運算元及所述第二運算元進行多個算術邏輯運算中的算術邏輯運算，

其中所述一或多個 DRAM 庫中的每一者包括：

DRAM 胞陣列，以多個頁配置，所述多個頁中的每一者包括多個 DRAM 胞，所述多個 DRAM 胞中的每一者儲存位元值；

列緩衝器，經組態以儲存所述多個頁中的打開頁的值；

輸入/輸出（IO）模組；以及

所述 IMC 模組，

其中所述 IMC 模組包括：

算術邏輯單元（ALU），經組態以自所述列緩衝器或所述 IO 模組接收運算元，以及基於所述運算元及所述多個算術邏輯運算中的所選算術邏輯運算來計算輸出；以及

結果暫存器，經組態以儲存由所述 ALU 計算出的所述輸出。

【請求項16】如請求項 15 所述的方法，其中所述多個資料佈局包括：

一個運算元（1OP）資料佈局，其中將所述第一運算元寫入至所述多個 DRAM 胞且將所述第二運算元自主機處理器直接供應至所述一或多個 DRAM 庫的所述 IMC 模組。

【請求項17】如請求項 16 所述的方法，其中所述 IMC 模組更包括運算元暫存器，且

其中所述記憶體控制器進一步經組態以：

將所述第一運算元的第一磚片儲存於所述運算元暫存器中；
以及

對儲存於所述運算元暫存器中的所述第一運算元及所述第二運算元的多個第二磚片中的每一者進行所述算術邏輯運算。

【請求項18】如請求項 15 所述的方法，其中將所述第一運算元劃

分成多個第一磚片且將所述第二運算元劃分成多個第二磚片，所述多個第一磚片中的每一者以及所述多個第二磚片中的每一者包括多個值，且

其中所述多個資料佈局包括相同頁（SP）資料佈局，其中所述記憶體控制器將所述多個第一磚片中的一或多者及所述多個第二磚片中的一或多者儲存於所述多個 DRAM 胞的相同頁中。

【請求項19】如請求項 18 所述的方法，其中所述 IMC 模組更包括運算元暫存器，且

其中所述記憶體控制器進一步經組態以：

將所述多個第一磚片中的一或多者中的第一磚片儲存於所述運算元暫存器中；以及

對儲存於所述運算元暫存器中的所述第一運算元及與所述第一磚片儲存於所述 DRAM 胞陣列的所述相同頁中的所述多個第二磚片中的一或多者中的每一者進行所述算術邏輯運算。

【請求項20】如請求項 19 所述的方法，其中所述一或多個 DRAM 庫中的至少一個的所述 IMC 模組更包括累加器，其中所述累加器包括經組態以儲存經累加值的累加器暫存器，其中所述累加器經組態以：

接收由所述 ALU 計算出的所述輸出；以及

用所述經累加值與所述輸出的總和更新所述累加器暫存器，
且

其中所述指令包括計算所述第一運算元與所述第二運算元的內積，其中所述多個第一磚片中的所述第一磚片儲存列資料，且其中所述多個第二磚片中的第二磚片包括行資料。

【請求項21】如請求項 19 所述的方法，其中所述第一磚片具有第一數目個值且所述多個第二磚片中的第二磚片具有第二數目個值，

其中所述一或多個 DRAM 庫中的至少一個的所述 IMC 模組包括輸出緩衝器，所述輸出緩衝器具有用於儲存大於或等於所述第一數目個值與所述第二數目個值的乘積的大小，且

其中所述指令包括計算所述第一運算元與所述第二運算元的外積，其中所述多個第一磚片中的所述第一磚片儲存列資料且所述多個第二磚片中的所述第二磚片包括行資料。

【請求項22】如請求項 19 所述的方法，其中所述第一磚片具有第一數目個值且所述多個第二磚片中的第二磚片具有第二數目個值，

其中所述一或多個 DRAM 庫中的至少一個的所述 IMC 模組包括輸出緩衝器，所述輸出緩衝器具有用於儲存大於或等於所述第一數目個值及所述第二數目個值中的更大者的大小，且

其中所述指令包括計算所述第一運算元與所述第二運算元的張量積，其中所述多個第一磚片中的所述第一磚片儲存列資料且所述第二磚片中的所述第二磚片包括行資料。

【請求項23】如請求項 15 所述的方法，其中將所述第一運算元劃分成多個第一磚片且將所述第二運算元劃分成多個第二磚片，所述多個第一磚片中的每一者以及所述多個第二磚片中的每一者包括多個值，且

其中所述多個資料佈局包括不同頁（DP）資料佈局，其中所述記憶體控制器儲存所述 DRAM 胞陣列的第一頁中的所述多個第

一磚片的子集及所述 DRAM 胞陣列的第二頁中的所述多個第二磚片的子集。

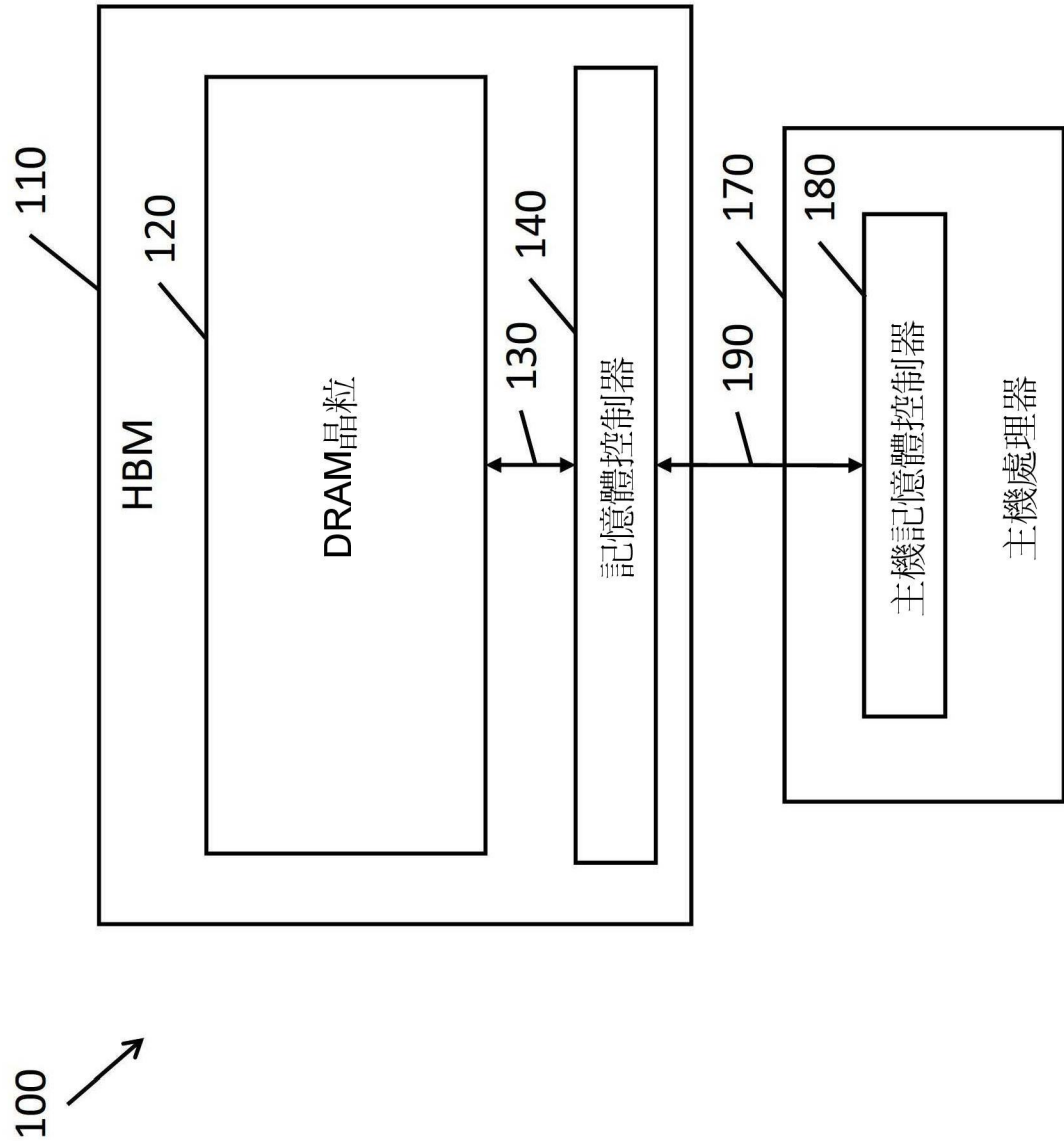
【請求項24】如請求項 23 所述的方法，其中所述 IMC 模組更包括運算元暫存器，且

其中所述記憶體控制器進一步經組態以：

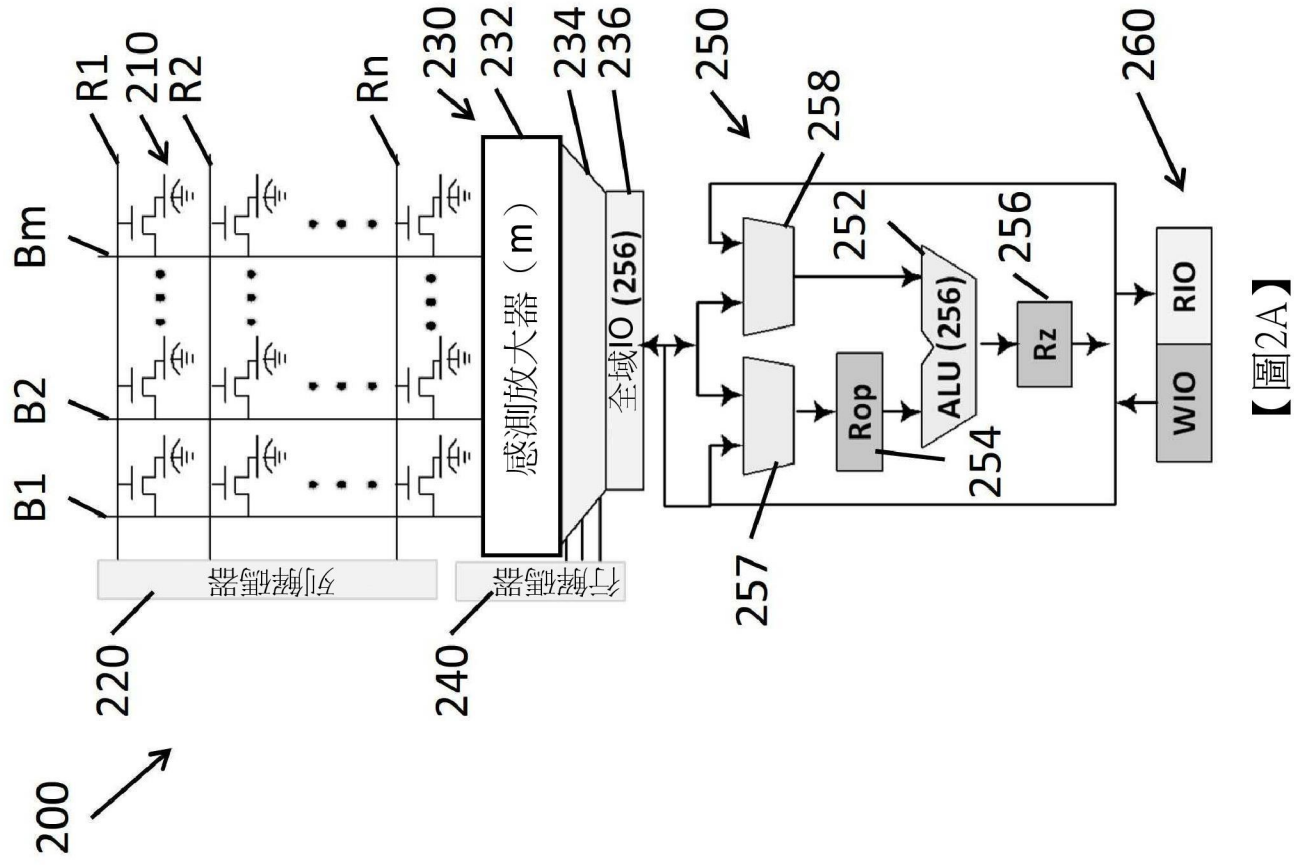
將來自所述第一頁的所述第一運算元的第一磚片儲存於所述運算元暫存器中；以及

對儲存於所述運算元暫存器中的所述第一運算元及來自所述第二頁的所述第二運算元的多個第二磚片中的每一者進行所述算術邏輯運算。

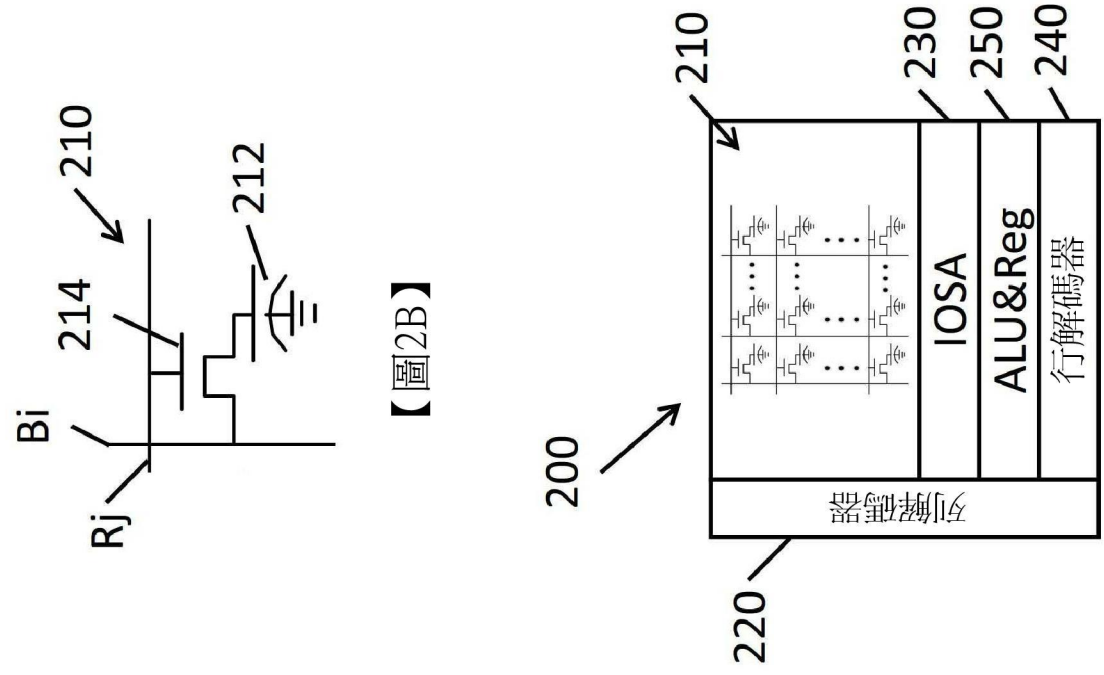
【發明圖式】



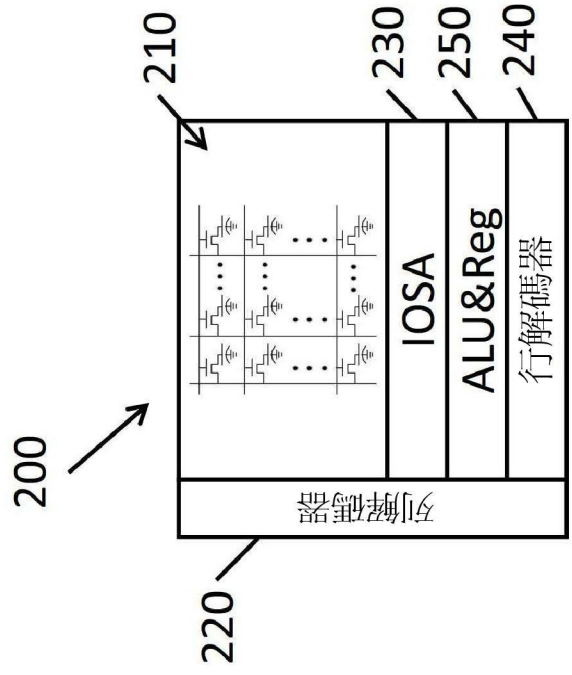
【圖1】



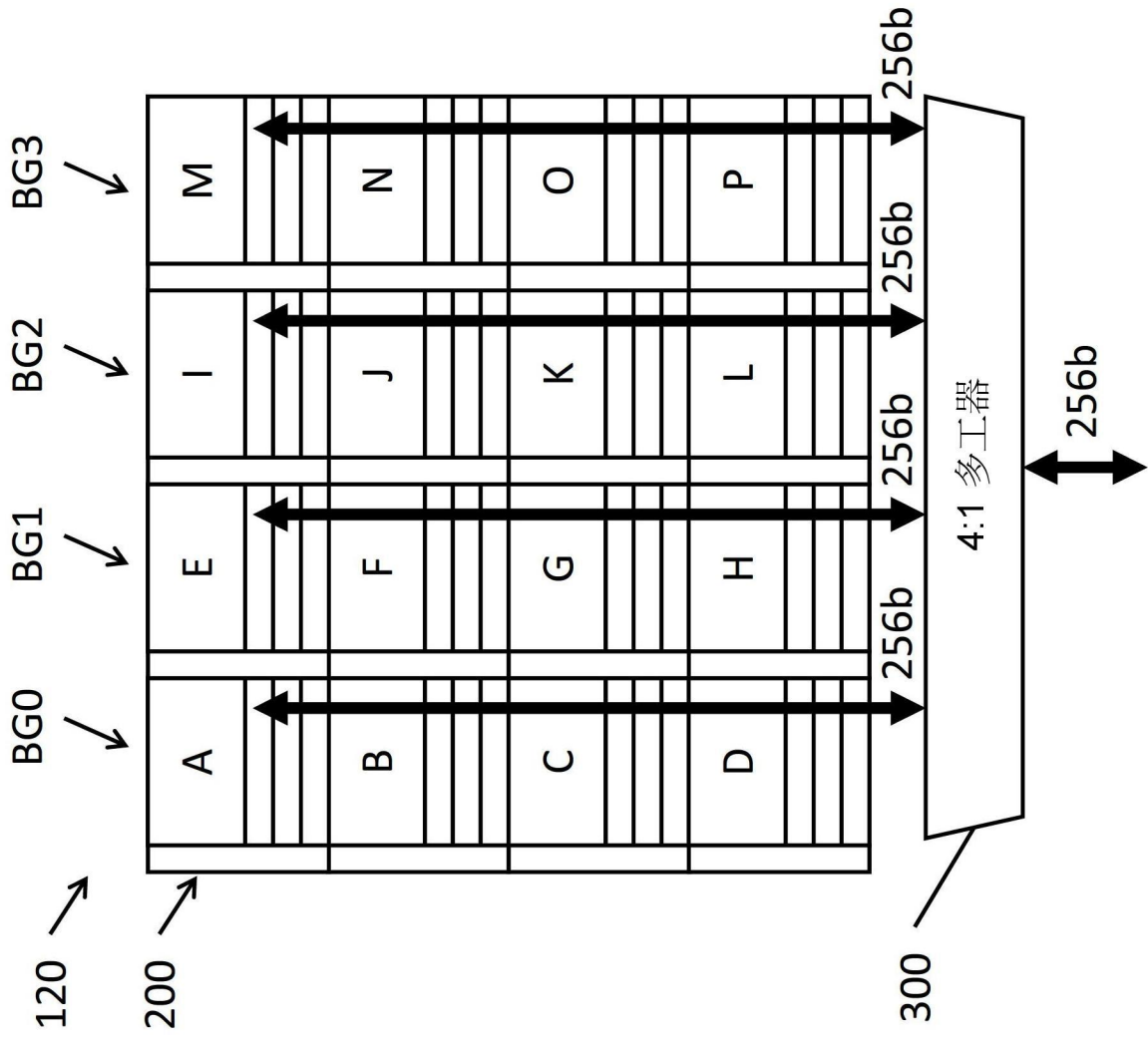
【圖2A】



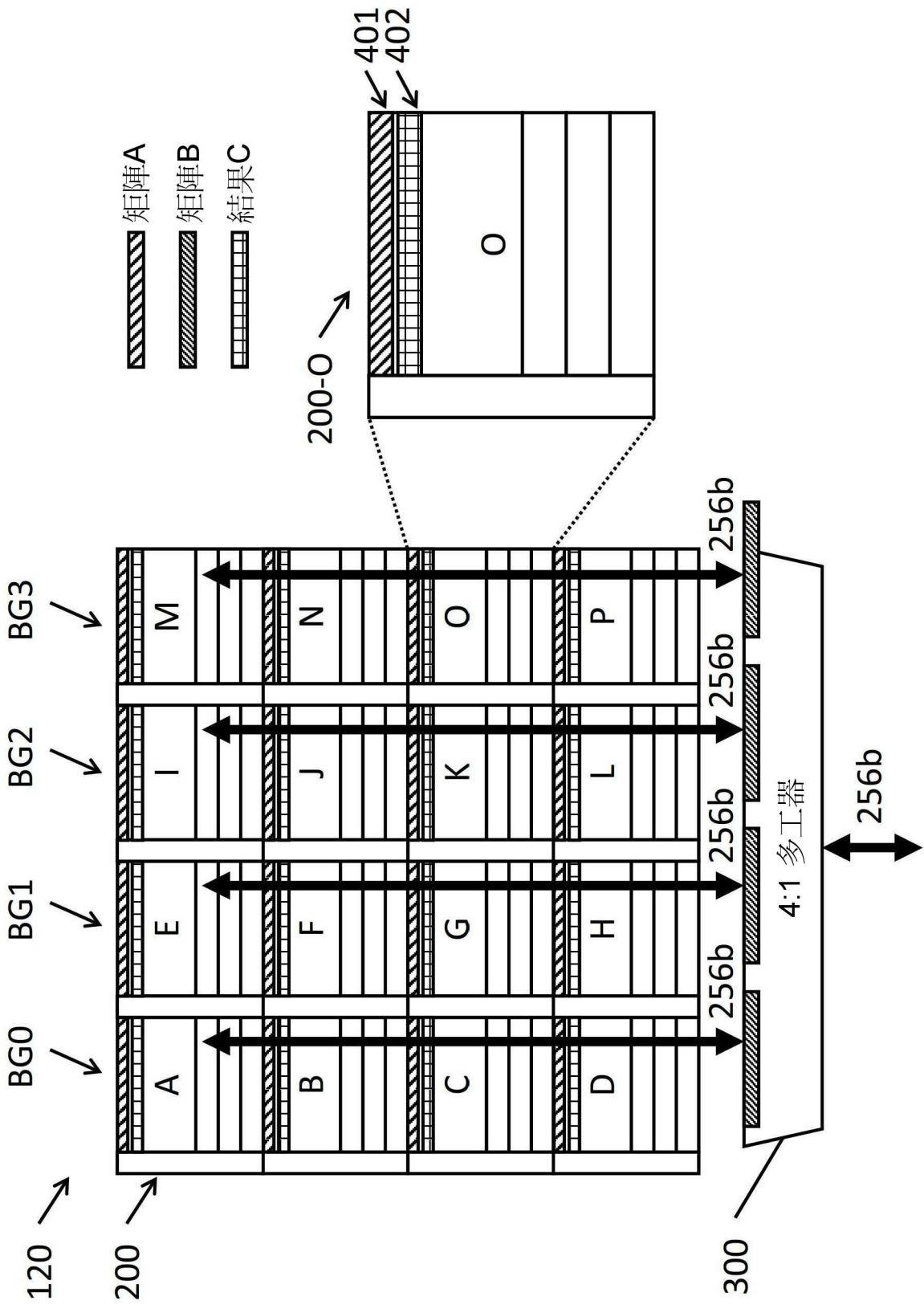
【圖2B】



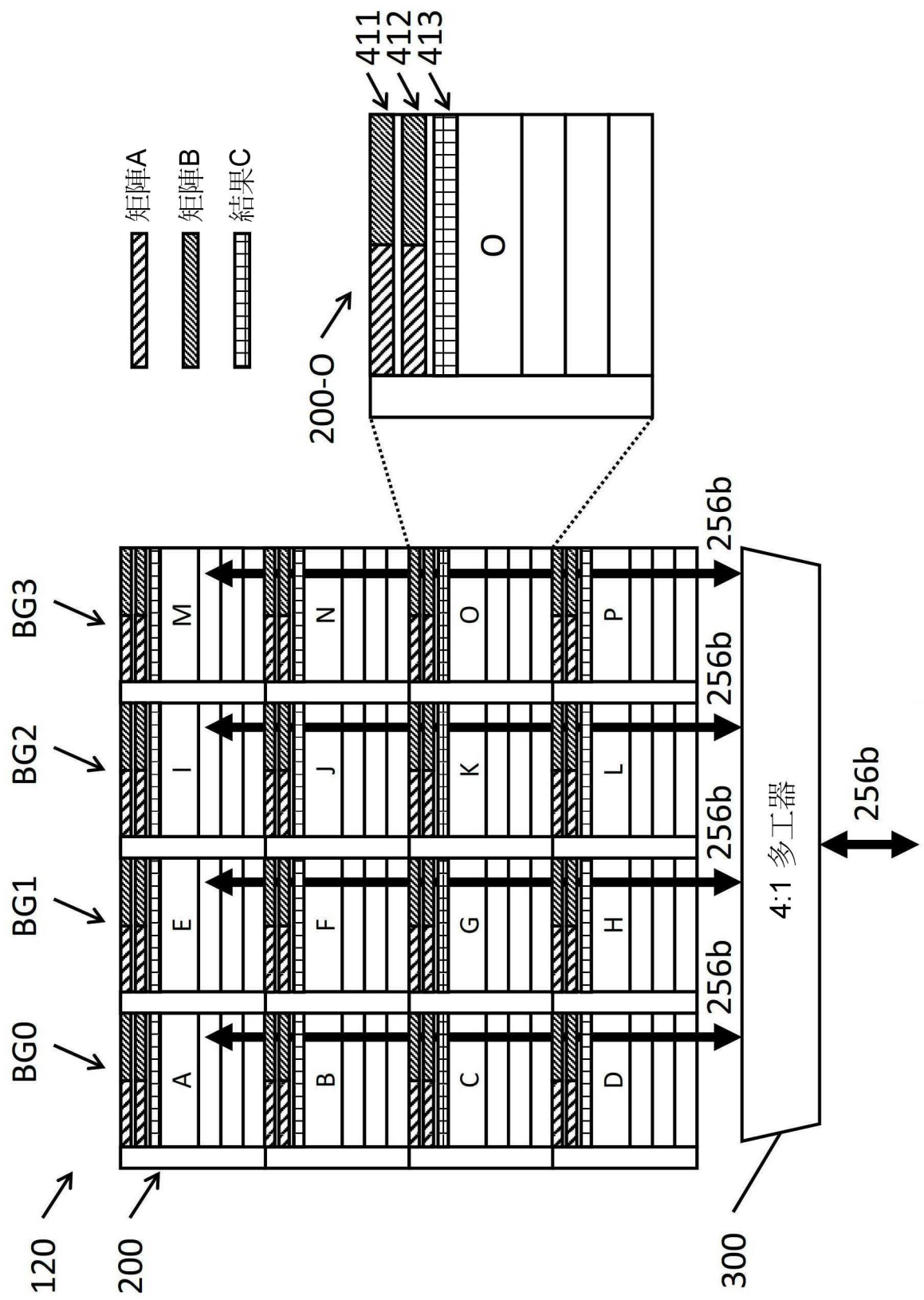
【圖2C】



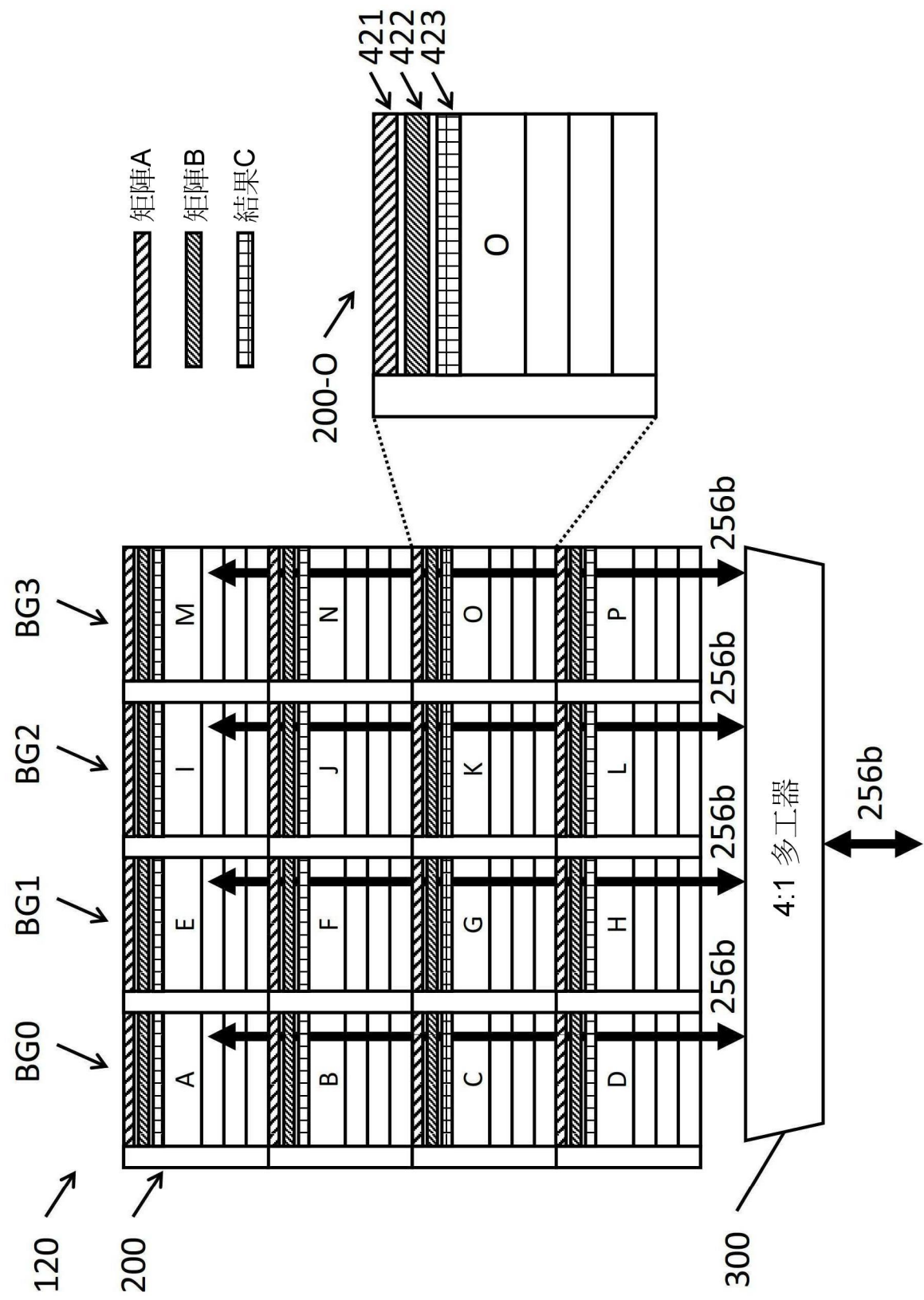
【圖3】



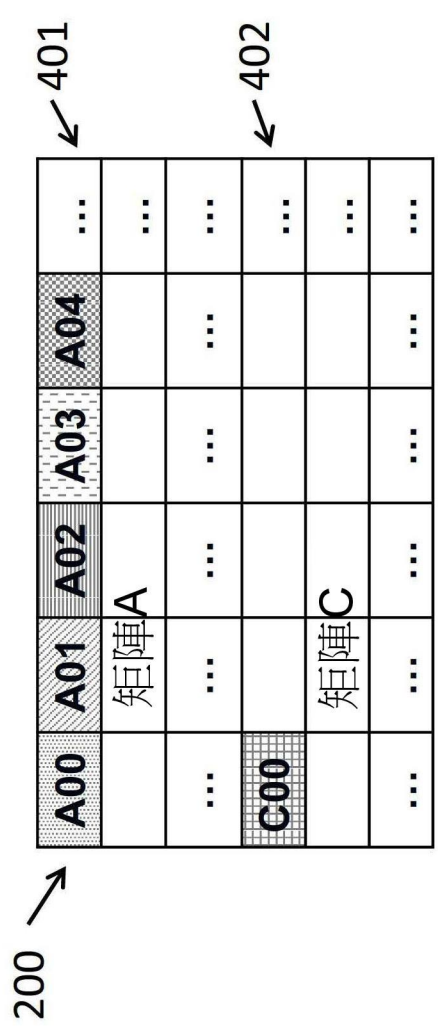
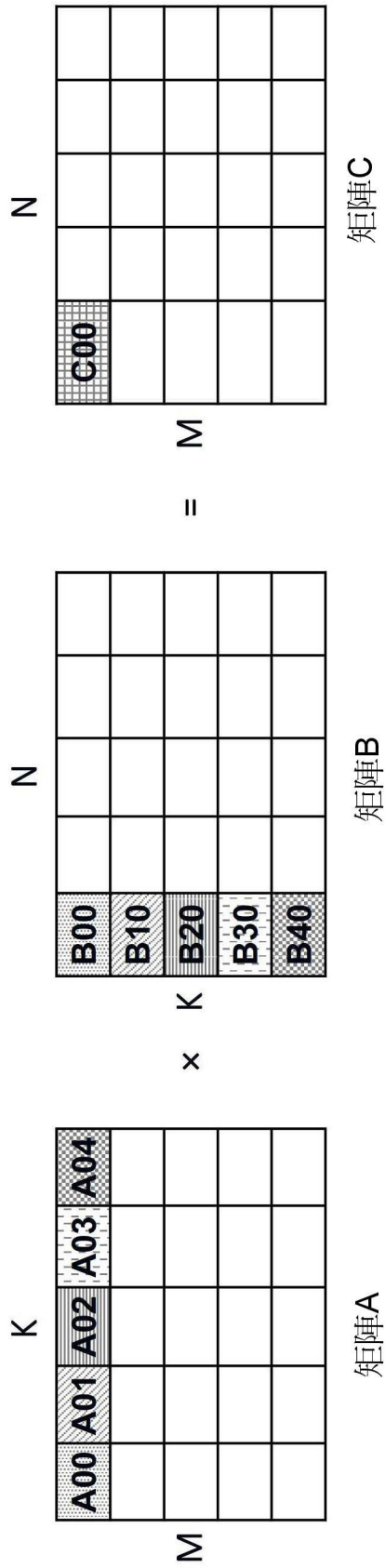
【圖4A】



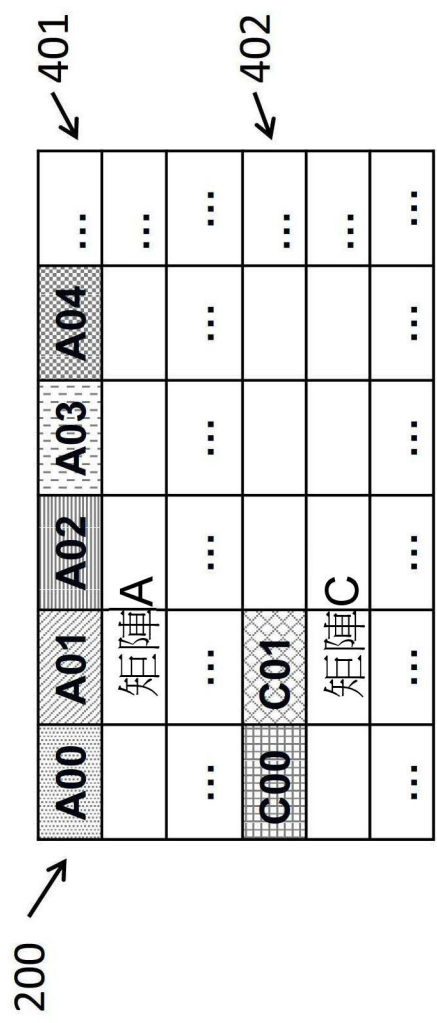
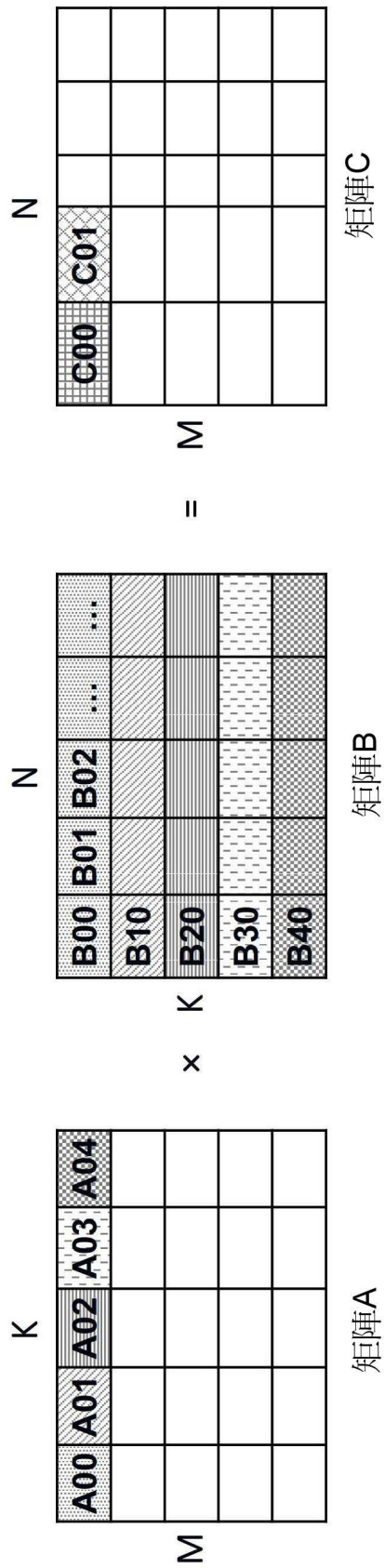
【圖4B】



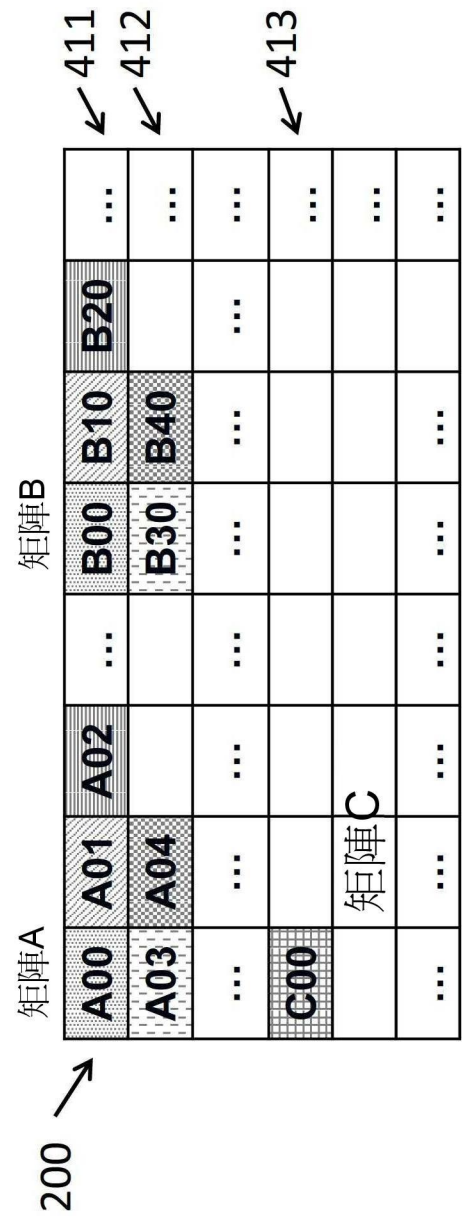
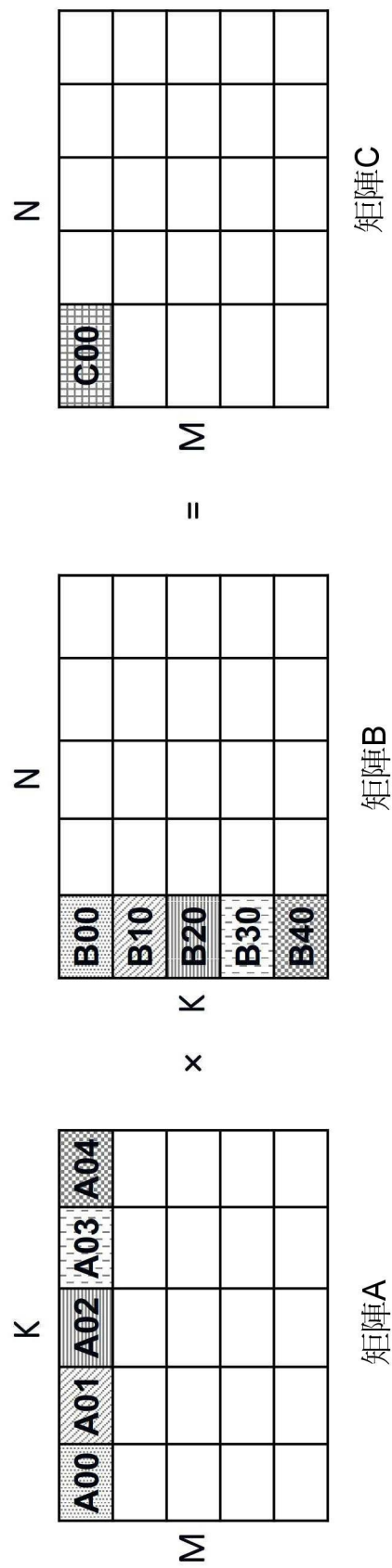
【圖4C】



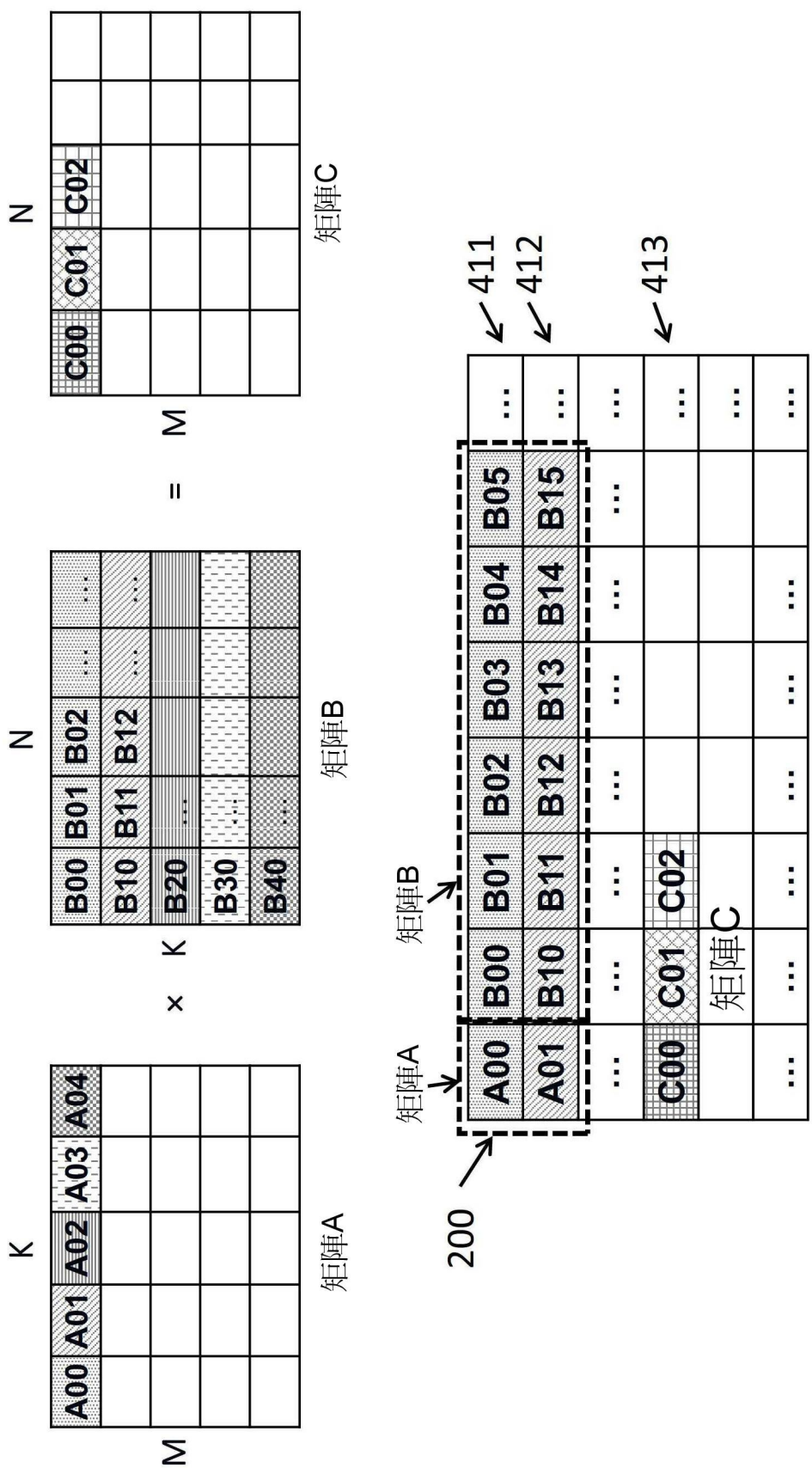
【圖5A】



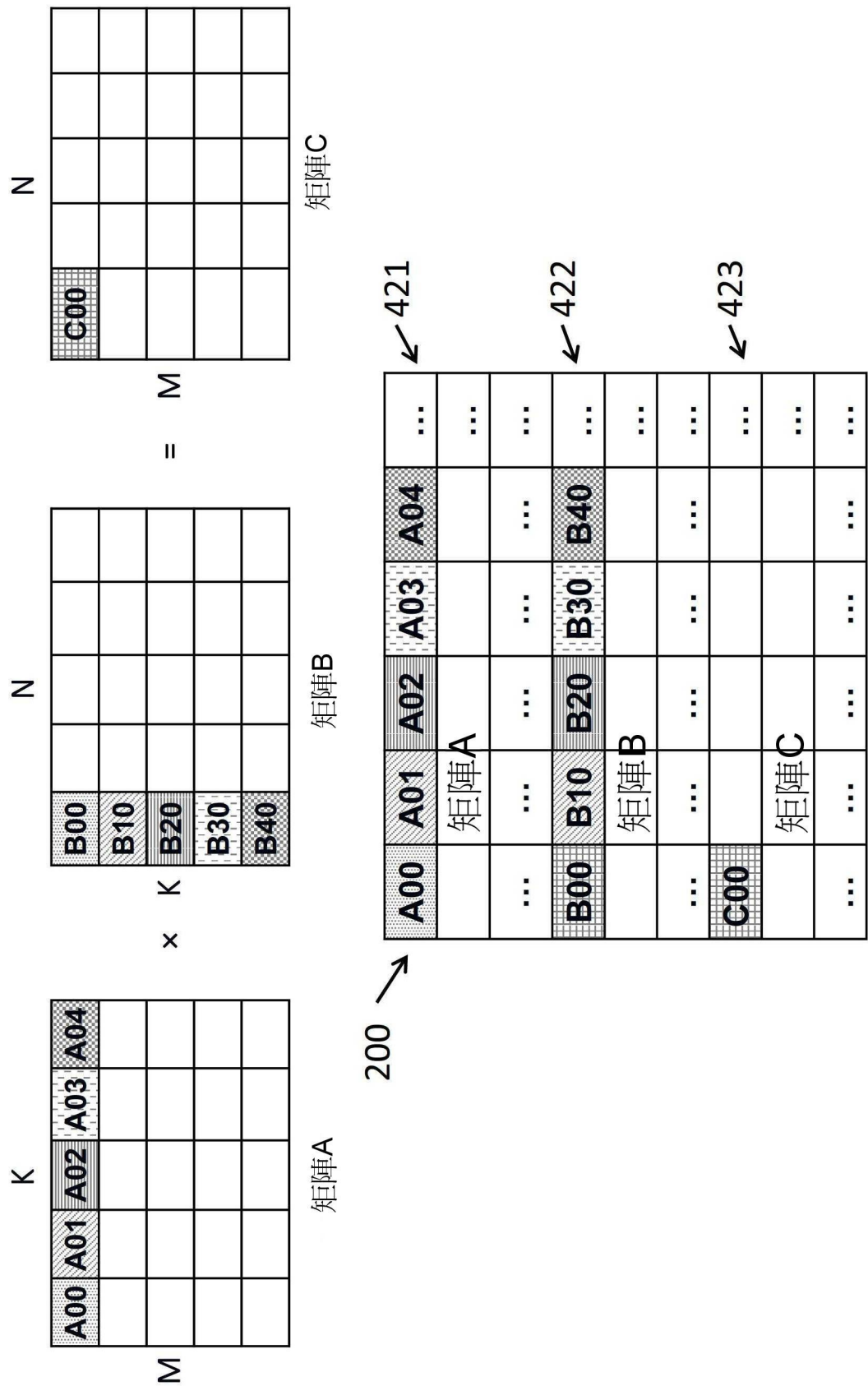
【圖5B】



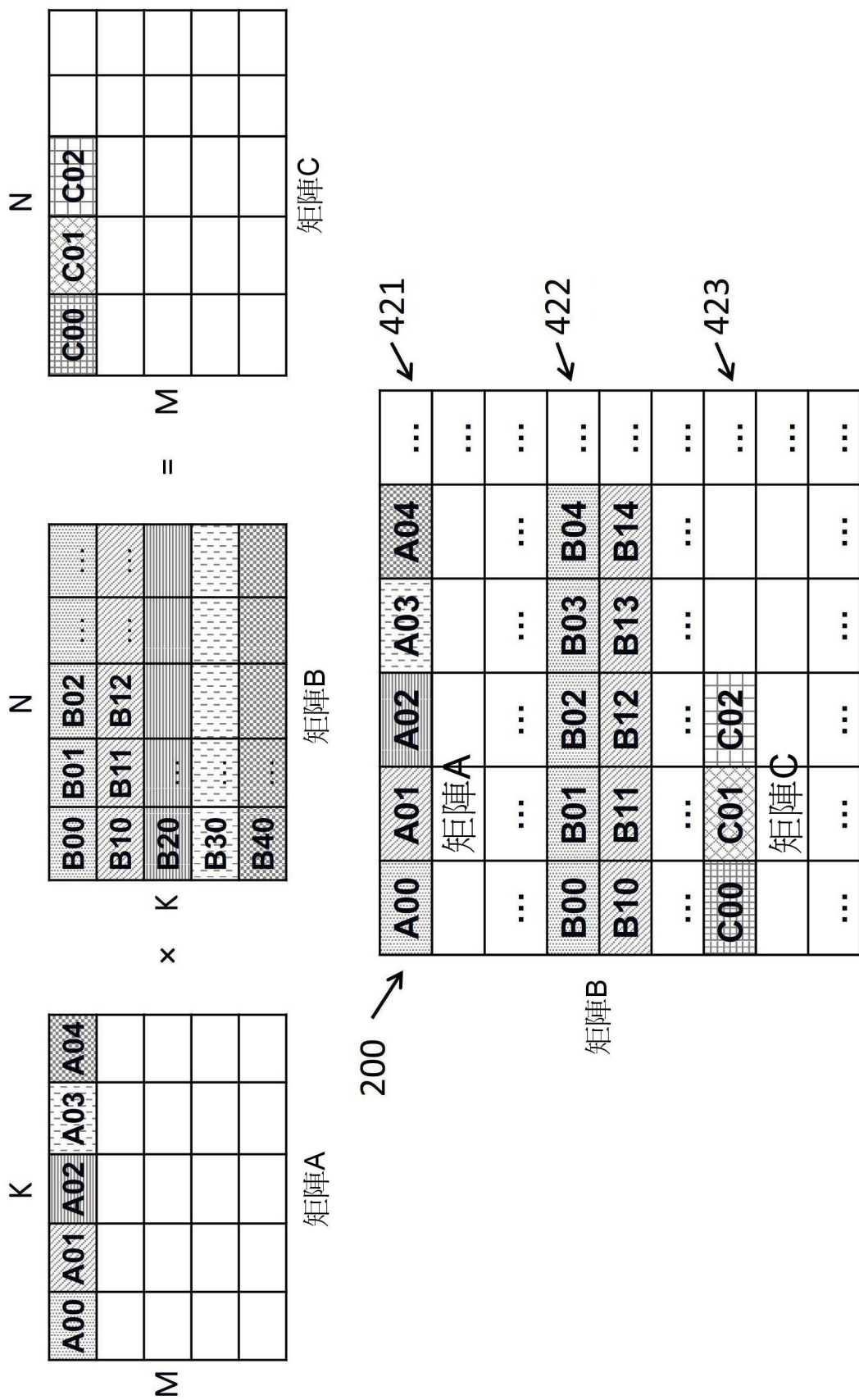
【圖6A】



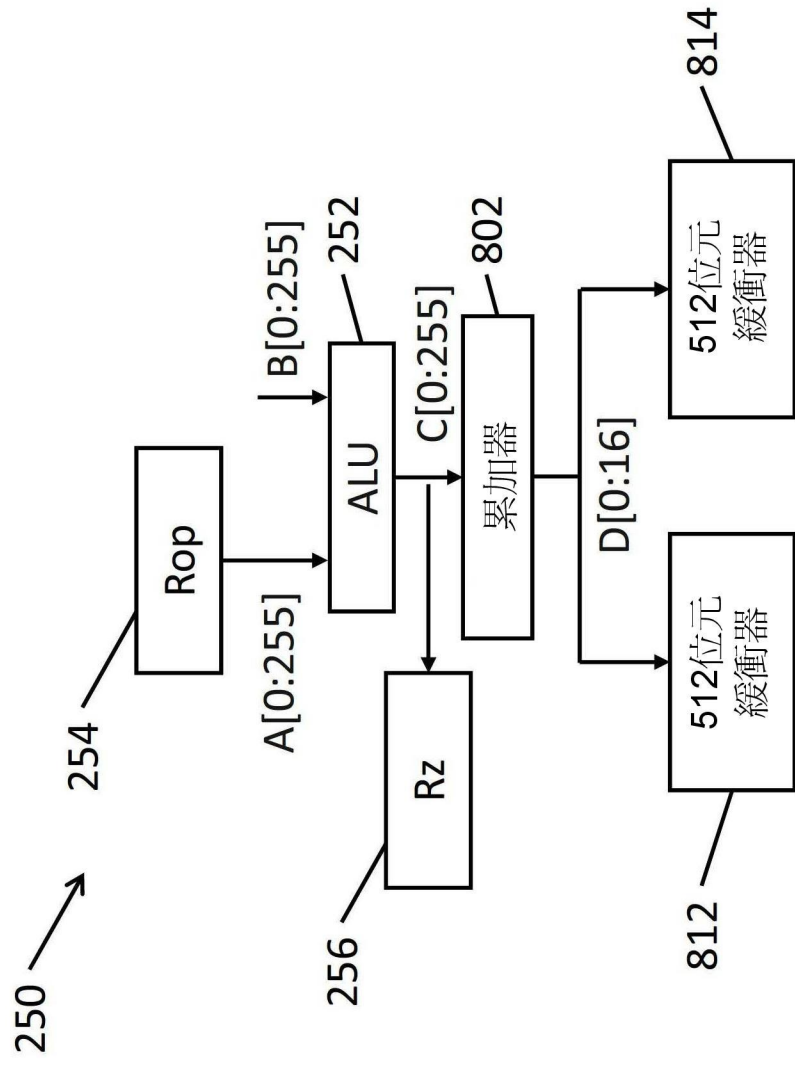
【圖6B】



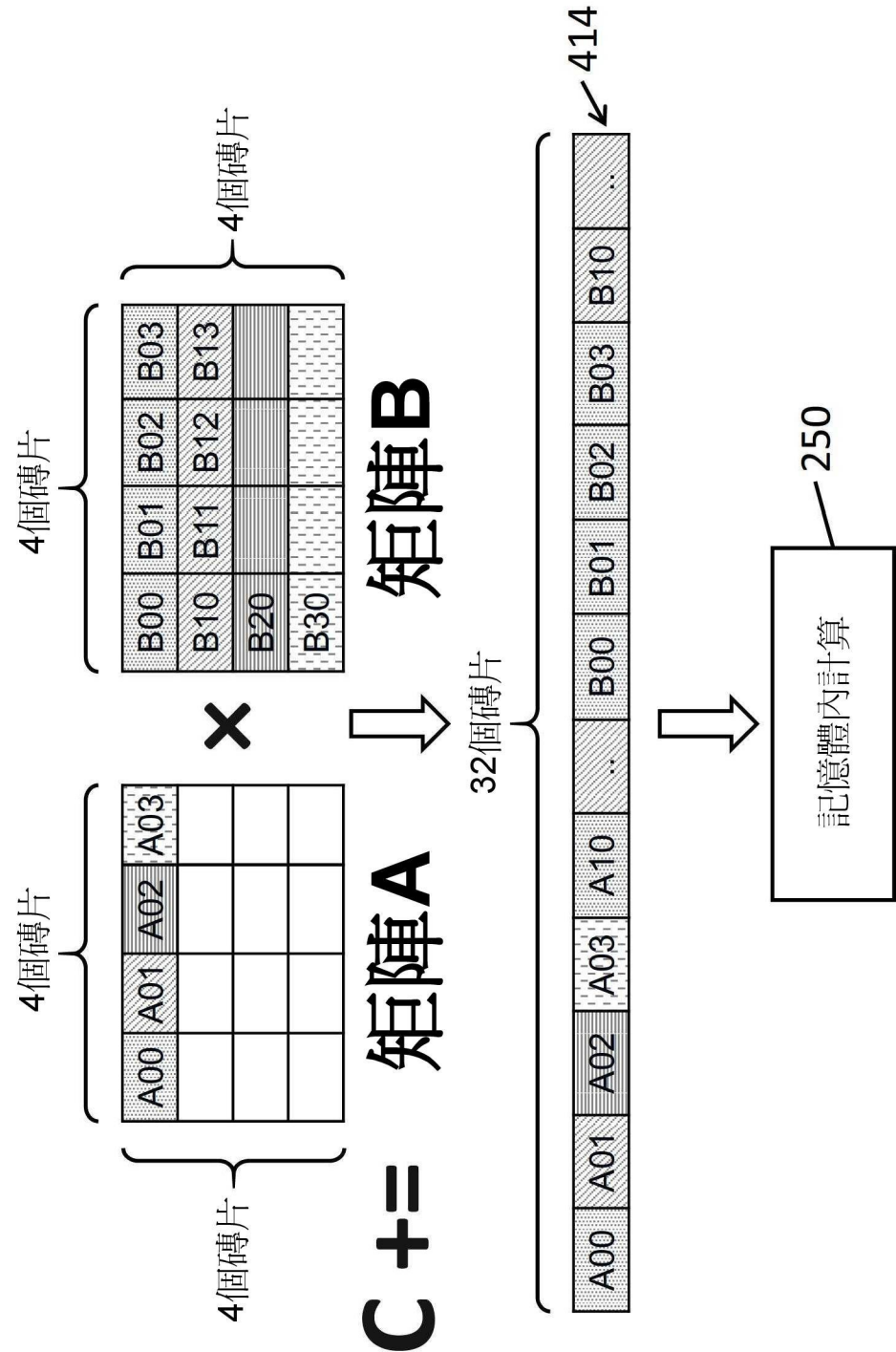
【圖7A】



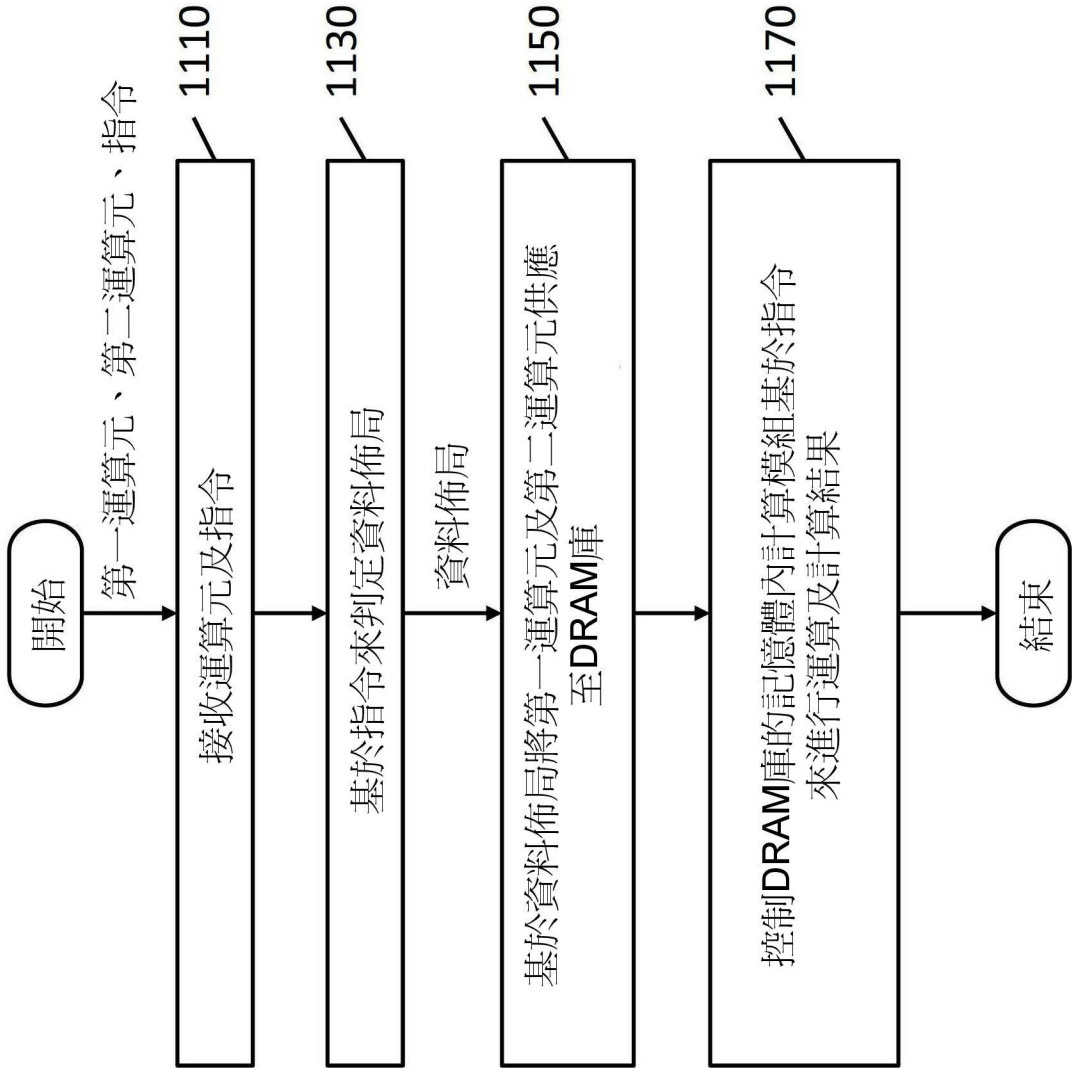
【圖7B】



【圖8】



【圖9】



【圖10】