I753728

(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 I753728 B**

(45)公告日： 中華民國 111 (2022) 年 01 月 21 日

(21)申請案號：109146644 (22)申請日： 中華民國 109 (2020) 年 12 月 29 日

(51)Int. Cl.： ***G06N3/063 (2006.01)*** ***G06N3/08 (2006.01)***

G06F13/36 (2006.01)

(71)申請人：財團法人工業技術研究院 (中華民國) INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE (TW)

新竹縣竹東鎮中興路四段 195 號

(72)發明人：陳耀華 CHEN, YAO-HUA (TW) ；嚴裕翔 YEN, YU-XIANG (TW) ；謝宛珊 HSIEH, WAN-SHAN (TW) ；黃稚存 HUANG, CHIH-TSUN (TW) ；盧俊銘 LU, JUIN-MING (TW) ；劉靖家 LIOU, JING-JIA (TW)

(74)代理人：許世正

(56)參考文獻：

TW I645301 US 10467501B2

審查人員：蔡夙勇

申請專利範圍項數：8 項 圖式數：4 共 19 頁

(54)名稱

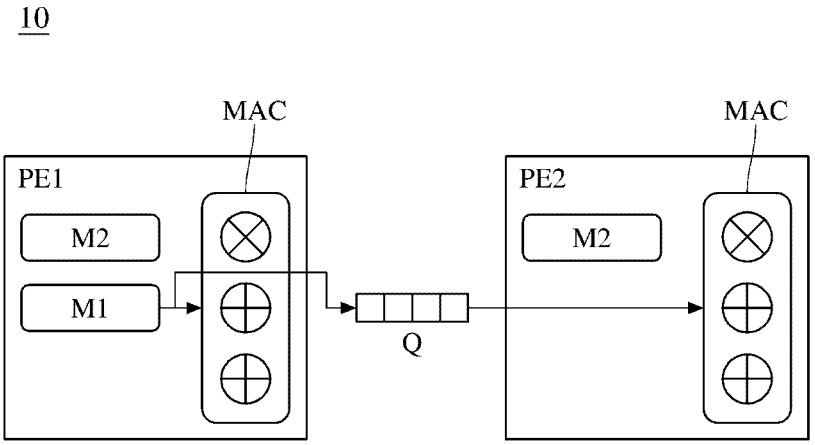
運算單元架構、運算單元叢集及卷積運算的執行方法

(57)摘要

適用於卷積運算的一種運算單元架構包括：多個運算單元以及一延遲佇列。運算單元中具有至少依據共用資料進行卷積運算的第一運算單元及第二運算單元。延遲佇列連接第一運算單元及第二運算單元。延遲佇列接收第一運算單元傳送的共用資料，並在接收共用資料且經過一延遲週期後將共用資料傳送至第二運算單元。

An architecture of processing elements adapted to a convolution operation comprises a plurality of processing elements and a delayed-queue. The plurality of processing elements has a first processing element and a second processing element which perform the convolution operation according to a shared data at least. The delayed-queue connects to the first processing element and the second processing element. The delayed-queue receives the shared data sent from the first processing element, and sends the shared data to the second processing element after receiving the shared data and passing a delayed period.

指定代表圖：



【圖1】

符號簡單說明：

10:運算單元架構

PE1:第一運算單元

PE2:第二運算單元

M1:第一儲存裝置

M2:第二儲存裝置

Q:延遲佇列

MAC:運算電路

專利案號：109146644

公告本

I753728

【發明摘要】

【中文發明名稱】　運算單元架構、運算單元叢集及卷積運算的執行方法

【英文發明名稱】　ARCHITECTURE AND CLUSTER OF PROCESSING ELEMENTS AND METHOD OF CONVOLUTION OPERATION

【中文】

適用於卷積運算的一種運算單元架構包括：多個運算單元以及一延遲佇列。運算單元中具有至少依據共用資料進行卷積運算的第一運算單元及第二運算單元。延遲佇列連接第一運算單元及第二運算單元。延遲佇列接收第一運算單元傳送的共用資料，並在接收共用資料且經過一延遲週期後將共用資料傳送至第二運算單元。

【英文】

An architecture of processing elements adapted to a convolution operation comprises a plurality of processing elements and a delayed-queue. The plurality of processing elements has a first processing element and a second processing element which perform the convolution operation according to a shared data at least. The delayed-queue connects to the first processing element and the second processing element. The delayed-queue receives the shared data sent from the first processing element, and sends the shared data to the second processing element after receiving the shared data and passing a delayed period.

【指定代表圖】

圖1

【代表圖之符號簡單說明】

10…運算單元架構

PE1…第一運算單元

PE2…第二運算單元

M1…第一儲存裝置

M2…第二儲存裝置

Q…延遲佇列

MAC…運算電路

【發明說明書】

【中文發明名稱】 運算單元架構、運算單元叢集及卷積運算的執行方法

【英文發明名稱】 ARCHITECTURE AND CLUSTER OF PROCESSING ELEMENTS AND METHOD OF CONVOLUTION OPERATION

【技術領域】

【0001】 本發明關於人工智慧，且涉及一種運行深度神經網路的人工智慧加速器。

【先前技術】

【0002】 深度神經網路（Deep Neural Network，DNN）近年來發展迅速。應用DNN進行影像辨識的精確度也逐漸提高，甚至比人類辨識得更為精準。為了配合DNN的計算需求，人工智慧加速器（即運行DNN模型的處理器）必須提升硬體效能。從穿戴裝置、行動通訊裝置以至於自駕車、雲端伺服器所用的人工智慧系統，其所需的運算量隨著裝置規模而指數性成長。

【0003】 一般而言，DNN專用的處理器須滿足計算力與輸入輸出頻寬兩方面的需求。增加運算單元（Processing Element，PE）的數量理論上可提升運算力，然而也需要一個適用於大量運算單元的資料網路架構才能將輸入資料即時地送入每一個運算單元。對於一個運算單元，其電路面積中佔最大比例部分的是儲存元件，其次才是控制邏輯與運算邏輯。考慮到大量運算單元所伴隨的功耗與電路面積，如何設計良好的資料傳輸路徑，藉此減少儲存元件的用量成為設計人工智慧加速器時的一個重要議題。

【發明內容】

【0004】 有鑑於此，本發明提出一種運算單元架構、運算單元叢集及卷積運算的執行方法，在保有人工智慧加速器原本的運算效能的同時減少所需的儲存空間，並且兼具延展性。

【0005】 依據本發明一實施例的一種運算單元架構，適用於一卷積運算，該架構包括：多個運算單元，該些運算單元中具有一第一運算單元及一第二運算單元，該第一運算單元及該第二運算單元至少依據一共用資料進行該卷積運算；以及一延遲佇列，連接該第一運算單元及該第二運算單元，該延遲佇列接收該第一運算單元傳送的該共用資料，並在接收該共用資料且經過一延遲週期後將該共用資料傳送至該第二運算單元。

【0006】 依據本發明一實施例的一種運算單元叢集，適用於一卷積運算，該叢集包括：一第一運算群，具有多個第一運算單元；一第二運算群，具有多個第二運算單元；一匯流排，連接該第一運算群及該第二運算群，該匯流排提供多個共用資料至每一該些第一運算單元；以及多個延遲佇列，該些延遲佇列中的一者連接該些第一運算單元中的一者及該些第二運算單元中的一者，該些延遲佇列中的另一者連接該些第二運算單元的二者，且每一該些延遲佇列傳遞該些共用資料中的一者；其中該第一運算群中的每一該些第一運算單元包括一儲存裝置，該儲存裝置用以儲存該些共用資料中對應的該者；且該第二運算群中的每一該些第二運算單元不包括該儲存裝置。

【0007】 依據本發明一實施例的一種卷積運算的執行方法，適用於本發明一實施例的運算單元架構，該方法包括：以該第一運算單元接收一輸入資料及

該共用資料並依據該輸入資料及該共用資料執行該卷積運算；以該第一運算單元傳送該共用資料至該延遲佇列； 以該延遲佇列等待該延遲週期；在該延遲佇列等待該延遲週期之後，以該延遲佇列傳送該共用資料至該第二運算單元；以及以該第二運算單元接收另一輸入資料，並依據該另一輸出資料及該共用資料進行該卷積運算。

【0008】 以上之關於本揭露內容之說明及以下之實施方式之說明係用以示範與解釋本發明之精神與原理，並且提供本發明之專利申請範圍更進一步之解釋。

【圖式簡單說明】

【0009】

圖1是本發明一實施例的運算單元架構的方塊圖；

圖2是本發明另一實施例的運算單元架構的方塊圖；

圖3是本發明一實施例的運算單元叢集的方塊圖；以及

圖4是本發明一實施例的卷積運算的執行方法的流程圖。

【實施方式】

【0010】 以下在實施方式中詳細敘述本發明之詳細特徵以及特點，其內容足以使任何熟習相關技藝者了解本發明之技術內容並據以實施，且根據本說明書所揭露之內容、申請專利範圍及圖式，任何熟習相關技藝者可輕易地理解本發明相關之構想及特點。以下之實施例係進一步詳細說明本發明之觀點，但非以任何觀點限制本發明之範疇。

【0011】本發明涉及人工智慧加速器中的處理單元陣列（Processing Element Array，PE Array）。處理單元陣列用於處理一或多個卷積（convolution）運算。處理單元陣列從總體緩衝器（global buffer，GLB）接收卷積運算時所需的輸入資料，例如輸入特徵圖（input feature map，ifmap）、卷積核（kernel map）以及部分和（partial sum）等。處理單元陣列中包含多個處理單元。一般而言，每個處理單元包含暫存記憶體（scratch pad memory，spad），用於暫存前述的輸入資料、乘積累加運算（Multiply Accumulate，MAC）器以及控制邏輯。

【0012】本發明提出的運算單元架構包括兩種運算單元：第一運算單元及第二運算單元，其中第一運算單元PE1的數量為1個，第二運算單元PE2的數量至少為1個以上。圖1及圖2分別繪示「一個第二運算單元」及「兩個第二運算單元」的二實施例。「兩個以上的第二運算單元」的實施例則可按照圖1及圖2自行推得。

【0013】圖1是本發明一實施例的運算單元架構的方塊圖。所述的運算單元架構適用於卷積運算，且包括多個運算單元以及一延遲佇列。圖1所示的運算單元架構10包括一個第一運算單元PE1、一個第二運算單元PE2以及一個延遲佇列Q。

【0014】第一運算單元PE1及第二運算單元PE2至少依據一共用資料進行卷積運算。在一實施例中，共用資料為卷積核或過濾器（filter）。第一運算單元PE1包括第一儲存裝置M1、第二儲存裝置M2及運算電路MAC。第二運算單元PE2的硬體結構類似於第一運算單元PE1，其差別在於第二運算單元PE2並沒有設置第一儲存裝置M1。實務上，第一儲存裝置M1用於暫存共用資料，例如卷積核或過濾器。第二儲存裝置M2用於暫存非共用資料，例如輸入特徵圖或部分和。運算電路MAC例如為乘積累加運算器。運算電路MAC依據取自第一儲存裝置M1

的卷積核、取自第二儲存裝置M2的輸入特徵圖、以及取自第二儲存裝置M2的部分和等資料進行卷積運算。卷積核屬於共用資料，輸入特徵圖及部分和屬於非共用資料。實務上，輸入特徵圖及部分和可分別儲存在兩個相異的儲存裝置，或是儲存在一個儲存裝置下的相異儲存空間，本發明對此不予限制。

【0015】 延遲佇列（delayed-control queue）Q連接第一運算單元PE1及第二運算單元PE2。延遲佇列Q用以接收第一運算單元PE1傳送的共用資料，並在接收共用資料且經過一延遲週期P後將共用資料傳送至第二運算單元PE2。實務上，延遲佇列Q具有先進先出（First In-First Out，FIFO）的資料結構。舉例說明如下，其中以T_k代表第k個單位時間；

在T_k時，第一運算單元PE1傳送共用資料F1至延遲佇列Q；

在T_{k+1}時，第一運算單元PE1傳送共用資料F2至延遲佇列Q；因此，

在第T_{k+P}時，第二運算單元PE2從延遲佇列Q接收到共用資料F1：且

在第T_{k+1+P}時，第二運算單元PE2從延遲佇列Q接收到共用資料F2。

【0016】 在本發明一實施例中，延遲週期P的數量級與卷積運算的步幅（stride）數值相同。舉例來說，若步幅為2，則延遲週期亦為2個單位時間。

【0017】 在本發明一實施例中，延遲佇列Q的儲存空間的大小（size）不小於卷積運算的步幅。舉例說明如下，若卷積運算的步幅為3，且第一運算單元PE1在T_k時取得共用資料F1並進行第一次卷積運算，則第一運算單元PE1將在T_{k+3}時取得共用資料F4並進行第二次卷積運算。然而，在T_{k+1}至T_{k+2}的期間，延遲佇列Q仍需要暫存來自第一運算單元PE1的共用資料F2及共用資料F3，且在T_{k+3}時，延遲佇列Q將共用資料F1傳送至第二運算裝置PE2。因此延遲佇列Q至少需要3個單位空間，用於儲存共用資料F1~F3。

【0018】 圖2是本發明另一實施例的運算單元架構10’的方塊圖。相較於前一實施例，此實施例的運算單元架構10’包括一個第一運算單元PE1、一個第二運算單元PE2a、另一個第二運算單元PE2b、一個延遲佇列Q1以及另一個延遲佇列Q2。第二運算單元PE2a及另一第二運算單元PE2b至少依據共用資料進行卷積運算。另一延遲佇列Q2連接第二運算單元PE2a及另一第二運算單元PE2b。此另一延遲佇列Q2接收第二運算單元PE2a傳送的共用資料，並在接收共用資料且經過延遲週期後將共用資料傳送至另一第二運算單元PE2b。實務上，可依據需求，自行增加串接在第一運算單元PE1後的多個第二運算單元PE2以及對應這些第二運算單元PE2的延遲佇列Q。由上述可知，運算單元架構10中的延遲佇列Q的數量與第二運算單元PE2的數量相同。

【0019】 圖3是本發明一實施例的運算單元叢集20的方塊圖。所述的運算單元叢集20適用於卷積運算，且包括第一運算群21、第二運算群22、匯流排23以及多個延遲佇列Q。第一運算群21及第二運算群22排列為M列N行的二維陣列。在M列中的每一者具有多個第一運算單元中的一個及多個第二運算單元中的$(N-1)$個。在圖3繪示的範例中，M＝3且N＝7，然而本發明並不限制M及N的數值大小。延遲佇列Q具有M組，這M組的每一者具有$(N-1)$個延遲佇列Q。

【0020】 第一運算群21具有M個第一運算單元PE1，第一運算群21中的每個第一運算單元PE1與前一實施例所述的第一運算單元PE1相同。第一運算單元PE1具有用以儲存共用資料的第一儲存裝置M1。

【0021】 第二運算群22具有$M \times (N-1)$個第二運算單元PE2。第二運算群22中的每個第二運算單元PE2不包括第一儲存裝置M1。

【0022】 匯流排23連接第一運算群21及第二運算群22。在本發明一實施例中，匯流排23至連接每一個第一運算單元PE1，且匯流排23連接至每一個第二運算單元PE2。匯流排23提供多個共用資料至每個第一運算單元PE1。匯流排23提供多個非共用資料至每個第一運算單元PE1及每個第二運算單元PE2。共用資料及非共用資料的來源例如為GLB。

【0023】 請參考圖3，運算單元叢集20的延遲佇列Q的數量有$M \times (N-1)$個，每個延遲佇列Q用以傳遞共用資料。

【0024】 這些延遲佇列Q中的一者連接這些第一運算單元PE1中的一者及這些第二運算單元PE2中的一者。這些延遲佇列Q中的另一者連接這些第二運算單元PE2的二者，且每個些延遲佇列Q傳遞這些共用資料中的一者。換個角度而言，第一運算群21中的每個第一運算單元PE1藉由一延遲佇列Q連接第二運算群22中的一個第二運算單元PE2。第二運算群22中位於同一列且相鄰二行的二個第二運算單元PE2透過該些延遲佇列中的一者彼此連接。

【0025】 圖4是本發明一實施例的卷積運算的執行方法的流程圖。圖4所示的卷積運算的執行方法適用於圖1所示的運算單元架構10、圖2所示的運算單元架構10’或圖3所示的運算單元叢集20。

【0026】 步驟S1為「第一運算單元PE1接收輸入資料及共用資料，並依據輸入資料及共用資料執行卷積運算」。輸入資料及共用資料例如由匯流排23傳送至第一運算單元PE1。

【0027】 步驟S2為「第一運算單元PE1傳送共用資料至第k個延遲佇列Q，其中k＝1」。k同時代表延遲佇列的編號及第二處理單元的編號。步驟S1及S2並不限制先後執行順序。步驟S1及S2可同時執行。

【0028】步驟S3為「第k個延遲佇列Q等待一延遲時間」。延遲時間的長度取決於卷積運算的步幅。

【0029】步驟S4為「第k個延遲佇列Q傳送共用資料至第k個第二運算單元PE2」。

【0030】步驟S5為「第k個第二運算單元PE2接收另一輸入資料，並依據另一輸出資料及共用資料進行卷積運算」。

【0031】步驟S6為「判斷第k個第二運算單元PE2是否為最後一個第二運算單元PE2」。若步驟S6的判斷結果為是，則結束本發明一實施例的卷積運算的執行方法。若步驟S6的判斷結果為否，則執行步驟S7。

【0032】步驟S7為「第k個第二運算單元PE2傳送共用資料至第k+1個延遲佇列Q」。步驟S7類似於步驟S2，步驟S7及S2皆為第一運算單元PE1或第二運算單元PE2將共用資料傳送至下一級的延遲佇列Q。步驟S8為「k=k+1」，即遞增k的值。依據運算單元架構10或10'中的第二運算單元PE2的數量，步驟S3~S8可能被重複執行複數次。

【0033】綜上所述，本發明提出的運算單元架構、運算單元叢集及卷積運算的執行方法藉由第二運算單元及延遲佇列的設計，可節省原本用於儲存共用資料的大量儲存裝置。當人工智慧加速器中屬於第二運算群的第二運算單元的數量愈多，應用本發明可節省的電路面積愈大，藉此也節省大量的功率消耗。

【0034】雖然本發明以前述之實施例揭露如上，然其並非用以限定本發明。在不脫離本發明之精神和範圍內，所為之更動與潤飾，均屬本發明之專利保護範圍。關於本發明所界定之保護範圍請參考所附之申請專利範

圍。

【符號說明】

【0035】

10、10'…運算單元架構

PE1…第一運算單元

PE2、PE2a、PE2b…第二運算單元

MAC…運算電路

M1…第一儲存裝置

M2…第二儲存裝置

Q、Q1、Q2…延遲佇列

20…運算單元叢集

21…第一運算群

22…第二運算群

23…資料匯流排

S1~S8…步驟

【發明申請專利範圍】

【請求項1】 一種運算單元架構，適用於一卷積運算，該架構包括：

多個運算單元，該些運算單元中具有一第一運算單元及一第二運算單元，該第一運算單元及該第二運算單元至少依據一共用資料進行該卷積運算；以及

一延遲佇列，連接該第一運算單元及該第二運算單元，該延遲佇列接收該第一運算單元傳送的該共用資料，並在接收該共用資料且經過一延遲週期後將該共用資料傳送至該第二運算單元；其中該延遲週期的數量級與該卷積運算的步幅數值相同。

【請求項2】 如請求項1的運算單元架構，其中該些運算單元中具有另一第二運算單元，該第二運算單元及該另一第二運算單元至少依據該共用資料進行該卷積運算；以及

該運算單元架構更包括另一延遲佇列，該另一延遲佇列連接該第二運算單元及該另一第二運算單元，該另一延遲佇列接收該第二運算單元傳送的該共用資料，並在接收該共用資料且經過該延遲週期後將該共用資料傳送至該另一第二運算單元。

【請求項3】 如請求項1的運算單元架構，其中該延遲佇列的儲存空間不小於該卷積運算的步幅。

【請求項4】 一種運算單元叢集，適用於一卷積運算，該叢集包括：

一第一運算群，具有多個第一運算單元；

一第二運算群，具有多個第二運算單元；

一匯流排，連接該第一運算群及該第二運算群，該匯流排提供多個共用資料至每一該些第一運算單元；以及

多個延遲佇列，該些延遲佇列中的一者連接該些第一運算單元中的一者及該些第二運算單元中的一者，該些延遲佇列中的另一者連接該些第二運算單元的二者，且每一該些延遲佇列傳遞該些共用資料中的一者；其中

該第一運算群中的每一該些第一運算單元包括一儲存裝置，該儲存裝置用以儲存該些共用資料中對應的該者；且

該第二運算群中的每一該些第二運算單元不包括該儲存裝置。

【請求項5】 如請求項4的運算單元叢集，其中該儲存裝置為第一儲存裝置，且每一該些第一運算單元及每一該些第二運算單元更包括：

一第二儲存裝置，用以儲存一非共用資料；以及

一運算電路，電性連接該第一儲存裝置及該第二儲存裝置，且該運算電路依據該些共用資料中對應的該者及該非共用資料執行該卷積運算。

【請求項6】 如請求項4的運算單元叢集，其中

該第一運算群及該第二運算群形成M列N行的二維陣列，該M列中的每一者具有該些第一運算單元中的一個及該些第二運算單元中的$(N-1)$個；

該些延遲佇列具有M組，該M組的每一者具有$(N-1)$個延遲佇列。

【請求項7】 一種運算單元架構的執行方法，適用於請求項1的運算單元架構，該方法包括：

以該第一運算單元接收一輸入資料及該共用資料並依據該輸入資料及該共用資料執行該卷積運算；

以該第一運算單元傳送該共用資料至該延遲佇列；

以該延遲佇列等待該延遲週期；

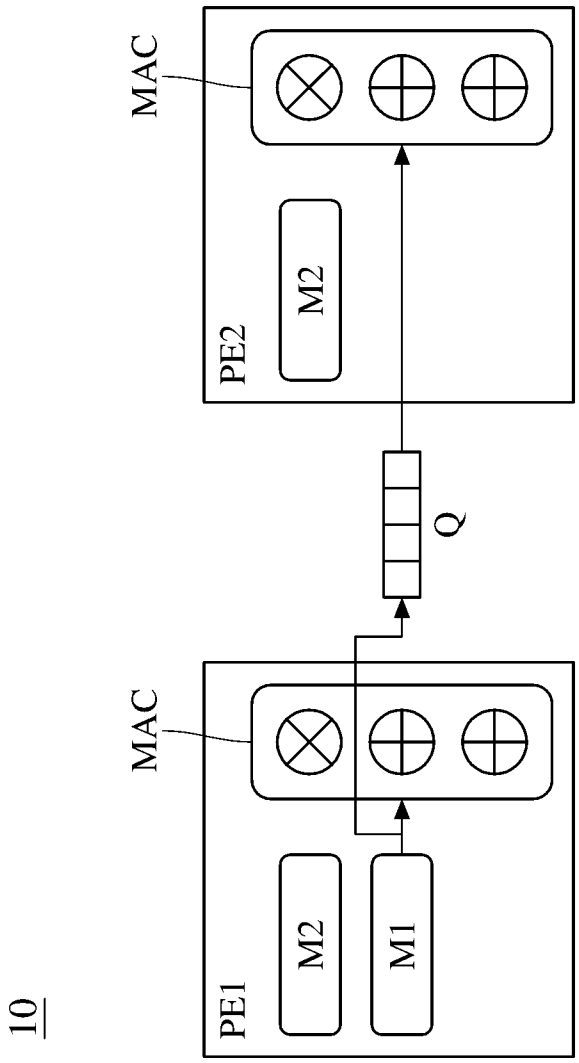
在該延遲佇列等待該延遲週期之後，以該延遲佇列傳送該共用資料至該第二運算單元；以及

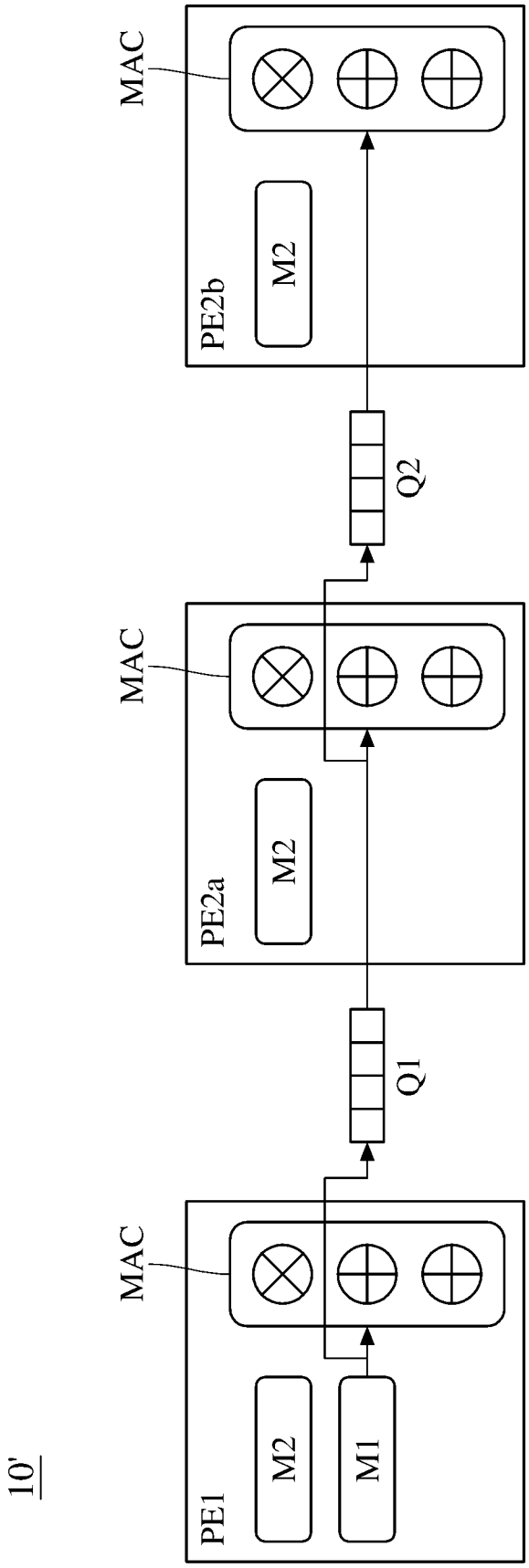
以該第二運算單元接收另一輸入資料，並依據該另一輸出資料及該共用資料進行該卷積運算。

【請求項8】如請求項7的卷積運算的執行方法，其中該共用資料為卷積核，且該輸入資料包括輸入特徵圖及部分和。

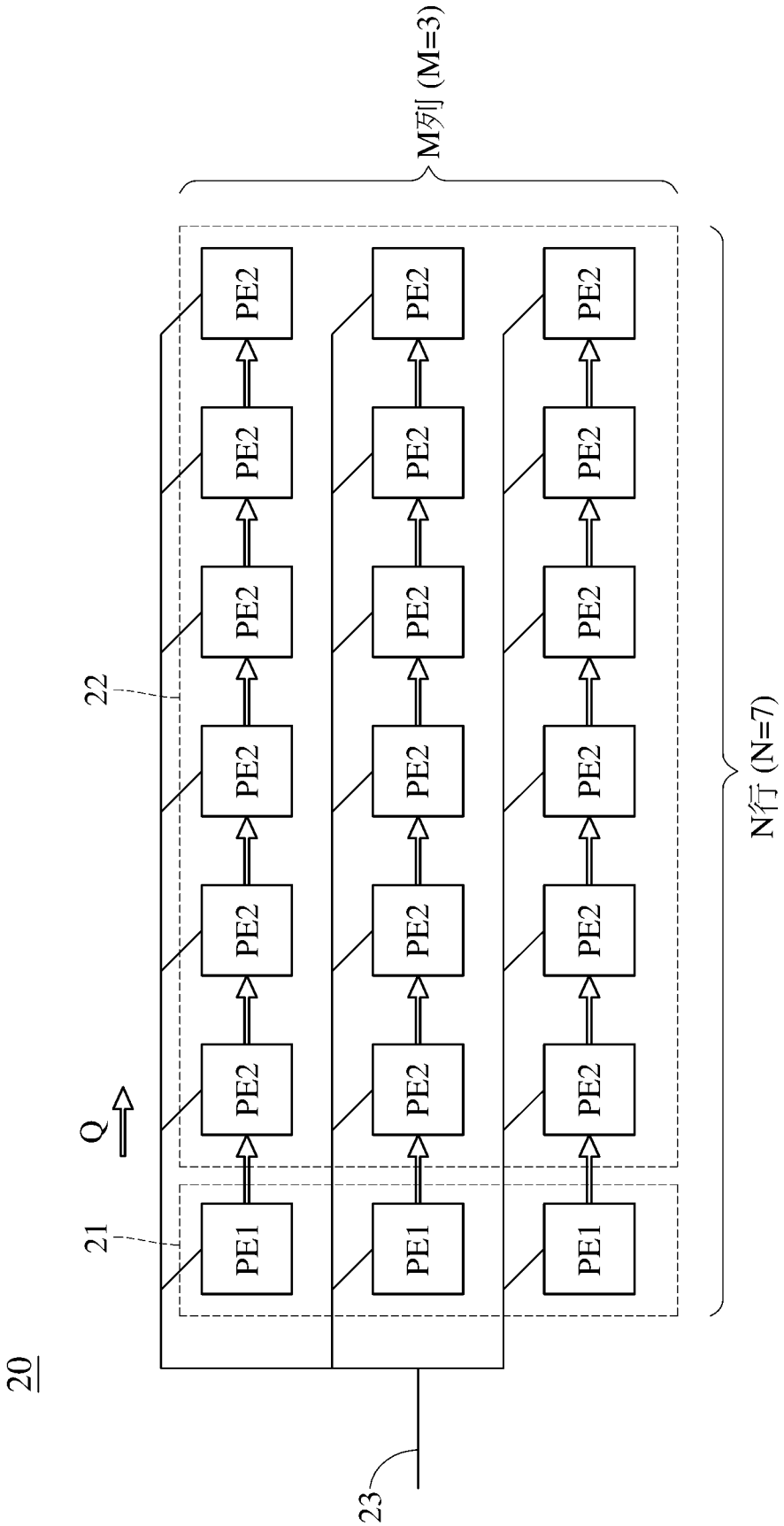
【發明圖式】

【圖1】

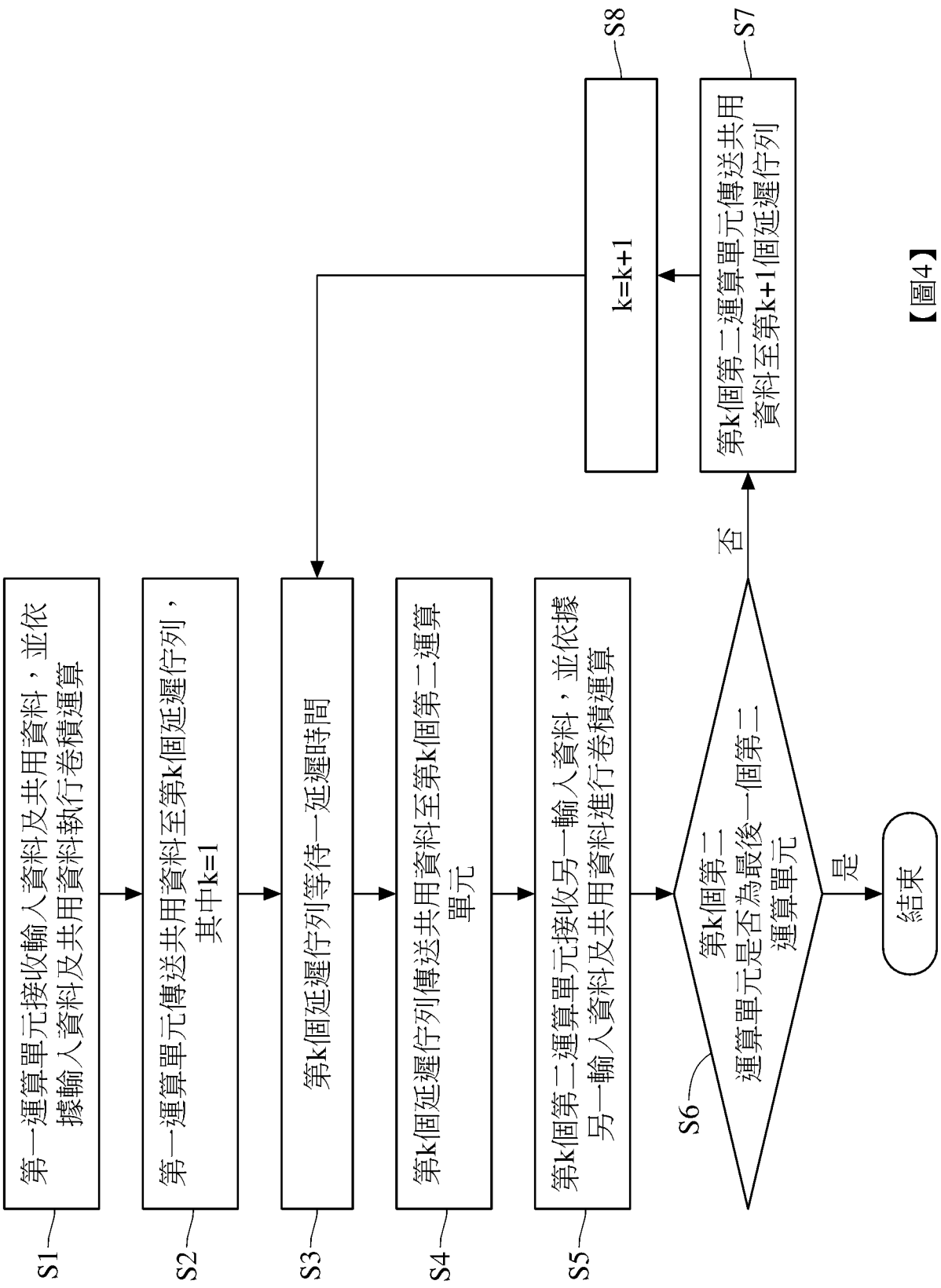




【圖2】



【圖3】


第一運算單元接收輸入資料及共用資料，並依據輸入資料及共用資料執行卷積運算
第一運算單元傳送共用資料至第k個延遲佇列，其中k=1
第k個延遲佇列等待一延遲時間
第k個延遲佇列傳送共用資料至第k個第二運算單元
第k個第二運算單元接收另一輸入資料，並依據另一輸入資料及共用資料進行卷積運算
第k個第二運算單元是否為最後一個第二運算單元
否
是
第k個第二運算單元傳送共用資料至第k+1個延遲佇列
k=k+1
結束
S1
S2
S3
S4
S5
S6
S7
S8

【圖4】