



(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 I868284 B

(45)公告日：中華民國 114 (2025) 年 01 月 01 日

(21)申請案號：109145287

(22)申請日：中華民國 109 (2020) 年 12 月 21 日

(51)Int. Cl. : G06F9/30 (2018.01)

G06F17/16 (2006.01)

G06T1/20 (2006.01)

G06N3/02 (2006.01)

(30)優先權：2020/05/05 印度

202041019059

2020/06/26 美國

16/913,800

(71)申請人：美商英特爾股份有限公司 (美國) INTEL CORPORATION (US)

美國

(72)發明人：麥猶蘭 沙布曼尼恩 MAIYURAN, SUBRAMANIAM (US)；帕拉 喬治 PARRA, JORGE (US)；帕爾 蘇普拉廷 PAL, SUPRATIM (IN)；蓋格 亞賈許 GARG, ASHUTOSH (IN)；馬瓦哈 蘇布拉 MARWAHA, SHUBRA (US)；古拉姆 錢德拉 GURRAM, CHANDRA (IN)；斯塔基 達林 STARKEY, DARIN (US)；波卡德格什 BORKAR, DURGESH (IN)；喬治 法吉斯 GEORGE, VARGHESE (US)

(74)代理人：林志剛

(56)參考文獻：

TW 201826122A

CN 110851779A

US 2014/0258689A1

US 2018/0189234A1

US 2019/0042542A1

US 2019/0294413A1

審查人員：詹効儒

申請專利範圍項數：20 項 圖式數：26 共 170 頁

(54)名稱

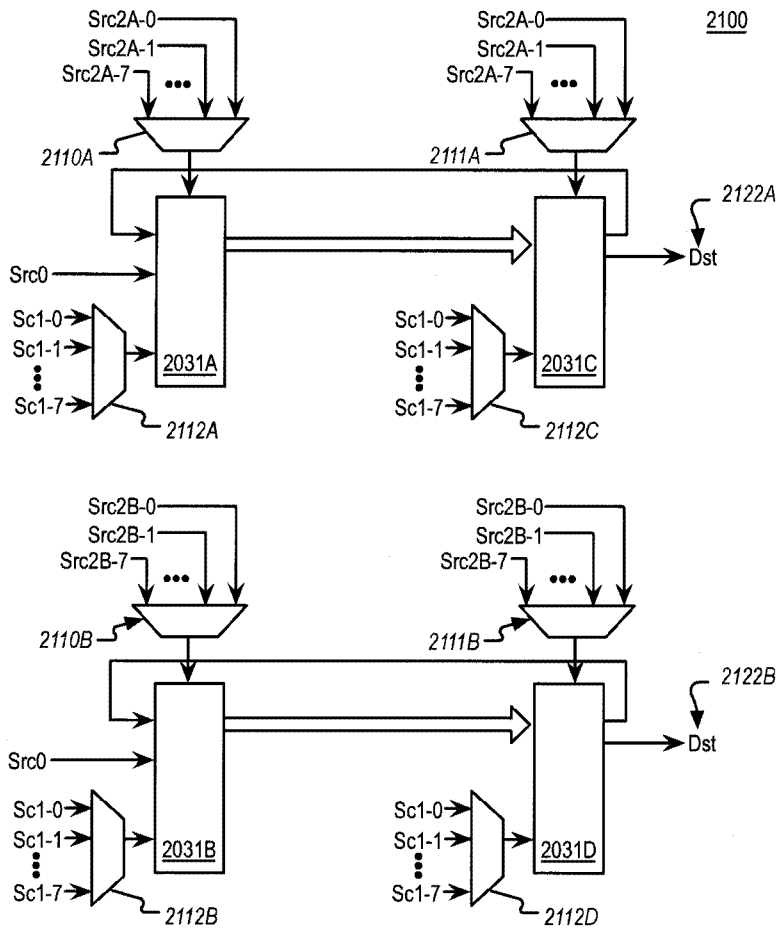
平行處理器、加速器裝置及執行平行處理之方法

(57)摘要

本文描述一種加速器裝置，其包含主機介面，與主機介面耦接的構造互連以及與構造互連耦接的一個或多個硬體磚，該一個或多個硬體磚包含稀疏矩陣乘法加速硬體，該稀疏矩陣乘法加速硬體包含具反饋輸入的脈動陣列。

Described herein is an accelerator device including a host interface, a fabric interconnect coupled with the host interface, and one or more hardware tiles coupled with the fabric interconnect, the one or more hardware tiles including sparse matrix multiply acceleration hardware including a systolic array with feedback inputs.

指定代表圖：



- 符號簡單說明：
- 2031A:處理元件
 - 2031B:處理元件
 - 2031C:處理元件
 - 2031D:處理元件
 - 2100:可擴縮的稀疏矩陣乘法加速器
 - 2110A:輸入選擇器
 - 2110B:輸入選擇器
 - 2111A:輸入選擇器
 - 2111B:輸入選擇器
 - 2112A:輸入選擇器
 - 2112B:輸入選擇器
 - 2112C:輸入選擇器
 - 2112D:輸入選擇器
 - 2122A:輸出邏輯
 - 2122B:輸出邏輯

【圖 21】



I868284

公告本

【發明摘要】

【中文發明名稱】

平行處理器、加速器裝置及執行平行處理之方法

【英文發明名稱】

PARALLEL PROCESSOR, ACCELERATOR DEVICE, AND
METHOD OF PERFORMING PARALLEL PROCESSING

【中文】

本文描述一種加速器裝置，其包含主機介面，與主機介面耦接的構造互連以及與構造互連耦接的一個或多個硬體磚，該一個或多個硬體磚包含稀疏矩陣乘法加速硬體，該稀疏矩陣乘法加速硬體包含具反饋輸入的脈動陣列。

【英文】

Described herein is an accelerator device including a host interface, a fabric interconnect coupled with the host interface, and one or more hardware tiles coupled with the fabric interconnect, the one or more hardware tiles including sparse matrix multiply acceleration hardware including a systolic array with feedback inputs.

【指定代表圖】第(21)圖。

【代表圖之符號簡單說明】

2031A:處理元件

2031B:處理元件

2031C:處理元件

2031D:處理元件

2100:可擴縮的稀疏矩陣乘法加速器

2110A:輸入選擇器

2110B:輸入選擇器

2111A:輸入選擇器

2111B:輸入選擇器

2112A:輸入選擇器

2112B:輸入選擇器

2112C:輸入選擇器

2112D:輸入選擇器

2122A:輸出邏輯

2122B:輸出邏輯

【特徵化學式】無

【發明說明書】

【中文發明名稱】

平行處理器、加速器裝置及執行平行處理之方法

【英文發明名稱】

PARALLEL PROCESSOR, ACCELERATOR DEVICE, AND
METHOD OF PERFORMING PARALLEL PROCESSING

【技術領域】

【0001】本案係主張於2020年5月5日申請的印度臨時專利申請案第202041019059號之優先權，其茲以引用方式併入本文中。

【0002】本發明係關於使用具反饋輸入之脈動陣列的可擴縮的稀疏矩陣乘法加速。

【先前技術】

【0003】用於機器學習工作負載之脈動矩陣乘法有相當大比例為零(稀疏-資料工作負載)。由於這些為零的乘法可以被跳過，因而改善整體性能。現有的脈動架構可為工作負載中的稀疏度提供支援，但此架構可能無法適度地擴縮。

【圖式簡單說明】

【0004】因此，可藉由參考以下詳細說明和所附圖式

的方式更具體地描述實施例來詳細地理解本實施例之特徵。然而，應注意的是，所附圖式僅描繪典型實施例，故不應被視為其範圍的限制。

【0005】[圖1]係根據實施例的處理系統之方塊圖；

【0006】[圖2A-2D]描繪由本文所述之實施例所提供的計算系統與圖形處理器；

【0007】[圖3A-3C]描繪由本文所述之實施例所提供之附加圖形處理器與計算加速器架構之方塊圖；

【0008】[圖4]係根據若干實施例之圖形處理器之圖形處理引擎410的方塊圖；

【0009】[圖5A-5B]描繪執行緒執行邏輯，其包含在根據本文所述之實施例的圖形處理器核心中採用的處理元件陣列；

【0010】[圖6]描繪根據實施例的附加執行單元；

【0011】[圖7]係根據若干實施例描繪圖形處理器指令格式之方塊圖；

【0012】[圖8]係根據另一實施例的圖形處理器之方塊圖；

【0013】[圖9A-9B]描繪根據若干實施例之圖形處理器命令格式與命令序列；

【0014】[圖10]描繪根據若干實施例針對資料處理系統之例示性圖形軟體架構；

【0015】[圖11A]係描繪根據一實施例之IP核心開發系統之方塊圖；

【0016】[圖 11B]描繪根據本文所述之若干實施例之積體電路封裝總成之橫截面側視圖；

【0017】[圖 11C]描繪封裝總成，其包含連接至基板之硬體邏輯小晶片之多個單元；

【0018】[圖 11D]描繪根據實施例之包含可互換小晶片的封裝總成；

【0019】[圖 12]係根據實施例的描繪例示性系統單晶片積體電路之方塊圖，其可使用一個或多個IP核心製造；

【0020】[圖 13A-13B]係根據本文所述之實施例描繪用於SoC內之例示性圖形處理器之方塊圖；

【0021】[圖 14]係根據實施例的資料處理系統之方塊圖；

【0022】[圖 15]描繪根據一實施例，由指令管線所執行之矩陣運算；

【0023】[圖 16]描繪乘法器/加法器電路的脈動陣列，其以管線的形式組織；

【0024】[圖 17A-17B]描繪使用四個深度的脈動陣列來計算八個脈動級(systolic stage)的等效陣列；

【0025】[圖 18A-18B]呈現脈動架構的時序圖；

【0026】[圖 19]描繪雙路徑矩陣乘法加速器，其中每一者路徑的深度為四級；

【0027】[圖 20]描繪四路徑矩陣乘法加速器，其中每一者路徑的深度為二級；

【0028】[圖 21]描繪使用具反饋輸入之脈動陣列的可

擴縮的稀疏矩陣乘法加速器；

【0029】[圖22]描繪包含稀疏資料之 Src2 輸入；

【0030】[圖23]描繪使用在每個級中具反饋輸入與輸出之脈動陣列的可擴縮的稀疏矩陣乘法加速器；

【0031】[圖24]描繪對本文所述之可擴縮的稀疏矩陣乘法加速器執行操作的方法；

【0032】[圖25]描繪使用稀疏 Src2 輸入矩陣執行矩陣乘法運算之方法；以及

【0033】[圖26]係根據實施例之包含圖形處理器的計算裝置之方塊圖。

【發明內容】與【實施方式】

【0034】本文描述使用具反饋輸入之脈動陣列實現可擴縮的稀疏矩陣乘法加速的裝置、系統、方法。

【0035】出於解釋的目的，闡述許多具體細節以供徹底地理解下述各種實施例。然而，本領域技術人員可清楚瞭解即使沒有這些具體細節亦可實施這些實施例。在其他情況下，習知的結構和裝置以方塊圖形式表示，以避免模糊基本原理，並提供對實施例更徹底的理解。雖然以下的一些實施例依據圖形處理器進行描述，但本文所述的技術和教示可應用於各種類型的電路或半導體裝置，包含通用處理裝置或圖形處理裝置。本文提及「一個實施例」或「一實施例」係表示所描述之與該實施例有關或相關聯的特定特徵、結構或特性可包含於此等實施例至少其中之一

中。然而，在說明書中各處出現的用語「在一個實施例中」並非必然皆指同一實施例。

【0036】在以下敘述和請求項中，可能使用術語「耦接」和「連接」及其衍生詞。應當理解，這些術語並非意旨彼此之間是同義字。「耦接」係用於表示兩個或多個元件可相互協作或互動，其彼此可直接或可不直接實體接觸或電性接觸。「已連接」係用於表示在彼此耦接的兩個或更多元件之間建立通訊。

【0037】以下的描述中，圖1至13A-13B提供結合或涉及各種實施例的示例性資料處理系統和圖形處理器邏輯的概述。圖14-26提供了各種實施例的具體細節。依據圖形處理器描述以下實施例的某些態樣，而其他態樣則依據諸如中央處理單元(CPU)之類的通用處理器描述。類似的技術和教示可應用於其他類型的電路或半導體裝置，包含但不限於多整合式核心處理器(many integrated core processor)、GPU群或現場可程式閘極陣列(FPGA)的一個或多個實例。一般而言，這些教示適用於任何操作或處理影像(例如，樣本、像素)、頂點資料或幾何資料的處理器或機器，或者執行用於機器學習和高性能計算應用的平行處理操作的任何處理器或機器。

系統概述

【0038】圖1係根據一實施例的處理系統100的方塊圖。系統100可用於單處理器桌上型系統、多處理器工作站系統或具有大量處理器102或處理器核心107的伺服器系

統中。在一實施例中，系統100是併入系統單晶片(SoC)積體電路內之處理平台，以用於具有對區域或廣域網路的有線或無線連接之行動、手持或嵌入式裝置(諸如物聯網(IoT)裝置等)中。

【0039】在一實施例中，系統100可包含、耦接、或整合於下述者中：基於伺服器之遊戲平台；遊戲主控台(包含遊戲與媒體主控台)；行動遊戲主控台(手持式遊戲主控台)；或線上遊戲主控台。在若干實施例中，系統100是行動電話、智慧型手機、平板計算裝置或行動網際網路連接裝置(諸如具有低內部儲存容量之筆記型電腦等)的一部分。處理系統100亦可包含、耦接、或整合於下述者中：穿戴式裝置(諸如智慧型手錶穿戴式裝置等)；具擴增實境(AR)或虛擬實境(VR)功能增強的智慧型眼鏡或服裝，可提供視覺、音訊或觸覺輸出，以補充真實世界中的視覺、音訊或觸覺體驗，或者另外提供文字、音訊、圖形、視訊、全像影像或視訊或觸覺反饋；其他擴增實境(AR)裝置；或其他虛擬實境(VR)裝置。若干實施例中，處理系統100包含或者是電視機或機上盒裝置的一部分。在一實施例中，系統100可包含、耦接、或整合於自動駕駛車輛，諸如公車、拖曳車(tractor trailer)、汽車、馬達或電動自行車、飛機或滑翔翼(或其任意組合)等。自動駕駛車輛可使用系統100以處理在車輛周遭感測到的環境。

【0040】在若干實施例中，一個或多個處理器102各包含一個或多個處理器核心107以處理指令，當該指令被

執行時將執行針對系統或使用者軟體之操作。在若干實施例中，一個或多個處理器核心 107 至少其中之一被配置以處理特定指令集 109。在若干實施例中，指令集 109 可促進複雜指令集計算 (CISC)、精簡指令集計算 (RISC) 或透過極長指令字 (VLIW) 之計算。一個或多個處理器核心 107 可處理不同指令集 109，其可包含用以促進其他指令集之仿真的指令。處理器核心 107 亦可包含其他處理裝置，諸如數位訊號處理器 (DSP) 等。

【0041】在若干實施例中，處理器 102 包含快取記憶體 104。取決於架構，處理器 102 可具有單一內部快取或多個層級的內部快取。在若干實施例中，快取記憶體是被共享於處理器 102 之各種組件間。在若干實施例中，處理器 102 亦使用外部快取 (例如，Level-3 (L3) 快取或最終階快取 (LLC) (未繪示出))，其可使用已知快取同調性技術而被共享於多個處理器核心 107 間。暫存器堆 106 可額外地包含於處理器 102 中，且可包含不同類型之暫存器以用於儲存不同類型之資料 (例如，整數暫存器、浮點暫存器、狀態暫存器、以及指令指標暫存器)。若干暫存器可為通用暫存器，而其他的暫存器可特定於處理器 102 之設計。

【0042】在若干實施例中，一個或多個處理器 102 與一個或多個介面匯流排 110 耦接，以在處理器 102 與系統 100 中其他組件之間傳送通訊訊號 (如，位址、資料或控制訊號)。在一實施例中，介面匯流排 110 可為諸如一版本之直接媒體介面 (DMI) 匯流排等的處理器匯流排。然而，處

理器匯流排並不限於 DMI 匯流排，且可包含一個或多個周邊組件互連匯流排(例如，PCI、PCI express)、記憶體匯流排、或其他類型之介面匯流排。在一實施例中，處理器 102 包含整合記憶體控制器 116 與平台控制器集線器 130。記憶體控制器 116 促進在記憶體裝置與系統 100 之其他組件之間的通訊，同時平台控制器集線器 (PCH) 130 透過區域 I/O 匯流排提供連接至 I/O 裝置。

【0043】記憶體裝置 120 可為動態隨機存取記憶體 (DRAM) 裝置、靜態隨機存取記憶體 (SRAM) 裝置、快閃記憶體裝置、相變記憶體裝置或若干其他具有適當性能可作為處理記憶體之記憶體裝置。在一實施例中，記憶體裝置 120 可作為系統記憶體運作而用於系統 100 來儲存資料 122 與指令 121，以供一個或多個處理器 102 於執行應用程式或處理時使用。記憶體控制器 116 亦與可選的外部圖形處理器 118 耦接，該外部圖形處理器 118 可與處理器 102 中一個或多個個圖形處理器 108 通訊以執行圖形與媒體操作。在若干實施例中，加速器 112 為協同處理器，其可輔助執行圖形、媒體及/或計算操作，且可被配置以執行專用的一組圖形、媒體、或計算操作。舉例而言，在一實施例中，加速器 112 為矩陣乘法加速器，用以最佳化機器學習或計算操作。在一實施例中，加速器 112 是光線追蹤加速器，其可用以與圖形處理器 108 合作執行光線追蹤操作。在一實施例中，外部加速器 119 可用以取代加速器 112 或與加速器 112 合作。

【0044】若干實施例中，顯示裝置111可連接至處理器102。顯示裝置111可為一個或多個內部顯示裝置(如同在行動電子裝置或筆記型電腦裝置中)或外部顯示裝置，其透過顯示介面(例如，DisplayPort等)附接。在一實施例中，顯示裝置111可為頭戴式顯示器(HMD)，諸如用於虛擬實境(VR)應用程式或擴增實境(AR)應用程式之立體顯示裝置。

【0045】在一些實施例中，平台控制器集線器130使周邊設備能夠透過高速I/O匯流排連接至記憶體裝置120和處理器102。I/O周邊設備包含，但不限於，音訊控制器146、網路控制器134、韌體介面128、無線收發器126、觸控感測器125、資料儲存裝置124(例如，非揮發性記憶體、揮發性記憶體、硬碟驅動器、快閃記憶體、NAND、3D NAND、3D XPoint等)。資料儲存裝置124可透過儲存介面(例如，SATA)或經由周邊匯流排(諸如，周邊組件互連匯流排(例如，PCI、PCI express))來連接。觸控感測器125可包含觸控螢幕感測器、壓力感測器或指紋感測器。無線收發器126可為Wi-Fi收發器、藍牙收發器或行動網路收發器(諸如3G、4G、5G或長程演進(LTE)收發器等)。韌體介面128賦能與系統韌體之通訊，且可例如為統一可延伸韌體介面(UEFI)。網路控制器134可賦能與有線網路之網路連接。在若干實施例中，高性能網路控制器(未繪示出)與介面匯流排110耦接。在一實施例中，音訊控制器146是多通道高解析音訊控制器。在一實施例中，系統100

包含可選的傳統 I/O 控制器 140，用於將舊有(例如，個人系統 2(PS/2))裝置耦接到系統。平台控制器集線器 130 亦可連接至一個或多個通用串列匯流排(USB)控制器 142 以連接輸入裝置(諸如，鍵盤與滑鼠 143 之組合、攝影機 144 或其他 USB 輸入裝置等)。

【0046】應瞭解，所繪示的系統 100 是例示性且非限制性，因為亦可使用其他類型具不同配置的資料處理系統。舉例而言，記憶體控制器 116 與平台控制器集線器 130 之實例可整合到一離散外部圖形處理器(諸如，外部圖形處理器 118)中。在一實施例中，平台控制器集線器 130 及/或記憶體控制器 116 可在一個或多個處理器 102 外部。舉例而言，系統 100 可包含外部記憶體控制器 116 與平台控制器集線器 130，其可被配置成在系統晶片組內與處理器 102 通訊的記憶體控制器集線器與周邊控制器集線器。

【0047】舉例而言，可使用其上設置有設計用於增加熱性能之 CPU、記憶體與其他組件等組件之電路板(「sled」，板)。在若干例示性中，處理組件(諸如處理器)是位於板頂側上，同時靠近位於該板底側上之記憶體(諸如 DIMM)。由於此種設計所提供之增強氣流之結果，造成組件可相較於典型系統在較高頻率與功率層級上運作，藉此增加效能。此外，板被配置以盲目搭接到機架中的電力與資料通訊電纜，藉此增強其可快速移除、升級、再安裝及/或替換之能力。類似地，位於板上之個別組件(諸如，處理器、加速器、記憶體及資料儲存驅動器)由於其彼此

間增加的間距而被配置成可輕易地被升級。在說明性實施例中，組件額外地包含硬體證實特徵，以證明其真實性。

【0048】資料中心可使用單一網路架構(「fabric」，構造)，其支援多種其他的網路架構(包含乙太網路與全路徑)。板可透過光纖而耦接到交換器，其比典型雙絞線電纜(例如，類別5、類別5e、類別6等)提供較高的頻寬與較低的延遲。由於該高頻寬、低延遲互連以及網路架構，因此資料中心在使用上可池化(pool)資源(諸如：記憶體、加速器(例如，GPU、圖形加速器、FPGA、ASIC、神經網路及/或人工智慧加速器等)以及實體上分開的資料儲存驅動器)，並根據需要將它們提供給計算資源(例如，處理器)，從而使計算資源能夠像其是本地(local)資源般一樣存取池化的資源。

【0049】電力供應或電源可提供電壓及/或電流至系統100或本文所述之任何組件或系統。在一範例中，電源包含AC至DC(交流至直流)配接器，其用以插入牆壁插座中。此種交流電可為可再生能源(例如太陽能)電源。在一範例中，電源包含DC電源，諸如外部AC至DC轉換器。在一範例中，電源或電力供應包含無線充電硬體，用以透過接近充電場來充電。在一範例中，電源可包含內部電池、交流電源、基於運動之電源、太陽能電源或燃料電池電源。

【0050】圖2A-2D描繪由本文所述實施例所提供的計算系統與圖形處理器。圖2A-2D之元件，其具有與本文其

他任意圖式之元件相同的參考編號(或名稱)者係可以類似於本文其他處所述者的任何方式來運作或運行，但不限於此。

【0051】圖2A是處理器200之實施例之方塊圖，處理器200具有一個或多個處理器核心202A-202N、整合記憶體控制器214、以及整合圖形處理器208。處理器200可包含附加核心，高達並包含由虛線框所表示之附加核心202N。各處理器核心202A-202N包含一個或多個內部快取單元204A-204N。在若干實施例中，各處理器核心亦可存取一個或多個共享快取單元206。內部快取單元204A-204N與共享快取單元206表示處理器200內快取記憶體階層。快取記憶體階層可包含各處理器核心內至少一層級之指令與資料快取，以及一個或多個層級之共享中階快取，諸如Level 2(L2)、Level 3(L3)、Level 4(L4)或其他層級之快取，其中在外部記憶體之前的最高層級快取被分類為LLC。在若干實施例中，快取同調性邏輯維持各種快取單元206與204A-204N之間的同調性。

【0052】在若干實施例中，處理器200亦可包含一組一個或多個匯流排控制器單元216與系統代理器核心210。一個或多個匯流排控制器單元216管理一組周邊匯流排，諸如一個或多個PCI或PCI express匯流排。系統代理器核心210對各種處理器組件提供管理功能。在若干實施例中，系統代理器核心210包含一個或多個整合記憶體控制器214，以管理對各種外部記憶體裝置(未繪示出)之存

取。

【0053】在若干實施例中，處理器核心 202A-202N 中一個或多個者包含支援同時多重執行緒。在此種實施例中，系統代理器核心 210 包含在多重執行緒處理期間用於協調與操作核心 202A-202N 之組件。系統代理器核心 210 可額外地包含功率控制單元(power control unit, PCU)，其包含邏輯與組件以調節處理器核心 202A-202N 與圖形處理器 208 之電力狀態(power state)。

【0054】在若干實施例中，處理器 200 額外地包含圖形處理器 208，以執行圖形處理操作。在若干實施例中，圖形處理器 208 與一組共享快取單元 206 及系統代理器核心 210 耦接，該系統代理器核心 210 包含一個或多個整合記憶體控制器 214。在若干實施例中，系統代理器核心 210 亦包含顯示控制器 211，以驅動圖形處理器對一個或多個耦接的顯示器進行輸出。在若干實施例中，顯示控制器 211 亦可為單獨的模組，其透過至少一互連而耦接到圖形處理器，或可整合於圖形處理器 208 中。

【0055】在若干實施例中，環式互連 212 是用以耦接該處理器 200 之內部組件。然而，亦可使用替代的互連單元，諸如點對點互連、交換式互連、或包含本領域中已知的其他技術。在若干實施例中，圖形處理器 208 透過 I/O 連接 213 來與環形互連 212 耦接。

【0056】例示性 I/O 連接 213 表示 I/O 互連之多種變體至少其中之一，包含封裝上 I/O 互連，其促進各種處理器

組件與高性能嵌入式記憶體模組 218(諸如 eDRAM 模組)之間的通訊。在若干實施例中，處理器核心 202A-202N 與圖形處理器 208 每一者可使用嵌入式記憶體模組 218 作為共享最終階快取。

【0057】在若干實施例中，處理器核心 202A-202N 是執行相同指令集架構之同質核心。在另一實施例中，處理器核心 202A-202N 就指令集架構 (ISA) 方面而論是異質的，其中一個或多個處理器核心 202A-202N 執行第一指令集，而其他核心至少其中之一執行該第一指令集之子集或不同的指令集。在一實施例中，處理器核心 202A-202N 就微架構方面而言是異質的，其中具有相對較高功耗之一個或多個核心與具有較低功耗之一個或多個電源核心相耦接。在一實施例中，處理器核心 202A-202N 就計算能力方面而言是異質的。額外地，處理器 200 可在一個或多個晶片上實現或被實現成除了其他組件外亦具有所描繪的組件之 SoC 積體電路。

【0058】圖 2B 係根據本文所述若干實施例的圖形處理器核心 219 之硬體邏輯的方塊圖。圖 2B 之元件，其具有與本文其他任意圖式之元件相同的參考編號(或名稱)者係可以類似於本文其他處所述者的方式來運作或運行，但不限於此。圖形處理器核心 219(有時被稱作核心切片)可為模組化圖形處理器中一個或多個圖形核心。圖形處理器核心 219 是一個圖形核心切片的示例，並且本文所述的圖形處理器可基於目標功率和性能範圍(performance envelope)包

含多個圖形核心切片。各圖形處理器核心219可包含與多個子核心221A-221F(亦被稱為子切片)耦接之固定功能區塊230，該子核心包含通用之模組化區塊與固定功能邏輯。

【0059】在若干實施例中，固定功能區塊230包含可被圖形處理器核心219中所有子核心共享之幾何/固定功能管線231，例如在較低性能及/或較低功率圖形處理器的實現中。在各種實施例中，幾何/固定功能管線231包含3D固定功能管線(例如，如圖3A與圖4中的3D管線312，如下述)、視訊前端單元、執行緒生成器與執行緒調度器、以及統一返回緩衝區管理器，其管理統一返回緩衝區(例如，圖4中的統一返回緩衝區418，如下述)。

【0060】在一實施例中，固定功能區塊230亦包含圖形SoC介面232、圖形微控制器233以及媒體管線234。圖形SoC介面232提供圖形處理器核心219與系統單晶片積體電路內其他處理器核心之間的介面。圖形微控制器233為可程式子處理器，其可配置以管理圖形處理器核心219中各種功能，包含執行緒調度、排程與搶占。媒體管線234(例如，圖3A與圖4之媒體管線316)包含用以促進多媒體資料之解碼、編碼、預處理及/或後處理之邏輯，該多媒體資料包含影像與視訊資料。透過向子核心221-221F內的計算或取樣邏輯之請求，媒體管線234實現媒體操作。

【0061】在一實施例中，SoC介面232賦能圖形處理器核心219與通用應用處理器核心(例如，CPU)及/或SoC內其

他組件進行通訊，該其他組件包含諸如共享最終階快取記憶體、系統RAM及/或嵌入式晶載或封裝上DRAM等記憶體階層元件。SoC介面232亦可賦能與SoC內的固定功能裝置(諸如攝影機成像管線)之通訊，並賦能使用及/或實現全局記憶體原子，該全局記憶體原子可在圖形處理器核心219與SoC內CPU之間被共享。SoC介面232亦可實現對圖形處理器核心219之功率管理控制，並賦能在圖形處理器核心219之時脈域與SoC內其他時脈域之間的介接。在一實施例中，SoC介面232賦能命令緩衝區從命令串流器和全局執行緒調度器的接收，該命令串流器和全局執行緒調度器被配置以向圖形處理器內的一個或多個圖形核心中的每一者提供命令和指令。當將執行媒體操作時，命令和指令可被調度到媒體管線234；或當將執行圖形處理操作時，命令和指令可被調度到幾何與固定功能管線(例如，幾何與固定功能管線231、幾何與固定功能管線237)。

【0062】圖形微控制器233可被配置以執行針對圖形處理器核心219之各種排程與管理任務。在一實施例中，圖形微控制器233可對在子核心221A-221F中的執行單元(EU)陣列222A-222F、224A-224F內的各種圖形平行引擎執行圖形及/或計算工作負載排程。在本排程模型中，執行於對SoC之CPU核心(包含圖形處理器核心219)上的主機軟體可提交多個圖形處理器門鈴其中之一的工作負載，其調用於適當的圖形引擎上的排程操作。排程操作包含判定接著要運行哪個工作負載、提交工作負載給命令串流器、搶

佔在引擎上運行之現存工作負載、監控工作負載之進度以及當工作負載完成時通知主機軟體。在一實施例中，圖形微控制器 233 亦可促進圖形處理器核心 219 之低電力或閒置狀態，提供圖形處理器核心 219 在低電力狀態轉換中獨立於作業系統及/或系統上圖形驅動器軟體之能力，以節省並復原圖形處理器核心 219 內的暫存器。

【0063】圖形處理器核心 219 可具有比所描繪的子核心 221A-221F 更多或更少之可高達 N 個的模組化子核心。針對各組 N 個子核心中，圖形處理器核心 219 亦可包含共享功能邏輯 235、共享及/或快取記憶體 236、幾何/固定功能管線 237 以及附加的固定功能邏輯 238，以加速各種圖形與計算處理操作。共享功能邏輯 235 可包含與圖 4 共享功能邏輯 420 相關聯之邏輯單元(取樣器、數學及/或執行緒間通訊邏輯)，其可被圖形處理器核心 219 內 N 個子核心每一者共享。共享及/或快取記憶體 236 可為圖形處理器核心 219 內一組 N 個子核心 221A-221F 的最終階快取，並且亦可作為可由多個子核心存取之共享記憶體。幾何/固定功能管線 237 可取代幾何/固定功能管線 231 包含於在固定功能區塊 230 中，且可包含相同或類似的邏輯單元。

【0064】在一實施例中，圖形處理器核心 219 包含附加的固定功能邏輯 238，其可包含各種被圖形處理器核心 219 所用之固定功能加速邏輯。在一實施例中，附加的固定功能邏輯 238 包含附加幾何管線，用於位置唯一著色(position-only shading)。在位置唯一著色中，存在兩個幾

何管線，即在幾何/固定功能管線 238、231 內之完整幾何管線以及剔除管線，該剔除管線可為包含於附加的固定功能邏輯 238 內之附加的幾何管線。在一實施例中，剔除管線為完整幾何管線之精簡版本。完整管線與剔除管線可執行相同應用程式之不同實例，各實例具有各自的上下文。位置唯一著色可隱藏丟棄的三角形之長剔除運行，從而使著色(shading)在某些情況下可更早完成。舉例而言，在一實施例中，附加的固定功能邏輯 238 內之剔除管線邏輯可與主應用程式平行執行位置著色器，並通常比完整管線較快產生關鍵結果，這是因為剔除管線僅提取並著色頂點之位置屬性而無執行光柵化(rasterization)及像素至訊框緩衝區之顯像(rendering)。剔除管線可使用產生之關鍵結果來計算所有三角形之能見度資訊，無需考慮此等三角形是否被剔除。完整管線(其在本實例中可被稱為重播管線)可使用能見度資訊以跳過剔除之三角形，以僅將最終通過光柵化階段之可見三角形進行著色。

【0065】在一實施例中，附加的固定功能邏輯 238 亦可包含機器學習加速邏輯，諸如固定功能矩陣乘法邏輯，以用於包含對機器學習訓練與推論之最佳化等之實現。

【0066】在各圖形子核心 221A-221F 內包含一組執行資源，其可用以執行圖形、媒體以及因應由圖形管線、媒體管線或著色器程式之請求而進行計算操作。圖形子核心 221A-221F 包含多 EU 陣列 222A-222F、224A-224F、執行緒調度與執行緒間通訊(TD/IC)邏輯 223A-223F、3D(例如，

紋理)取樣器 225A-225F、媒體取樣器 226A-226F、著色器處理器 227A-227F、以及共享區域記憶體 (SLM)228A-228F。EU陣列 222A-222F、224A-224F每一者包含多個執行單元，其為能執行浮點與整數/固定點邏輯操作之通用圖形處理單元，以服務於圖形、媒體或計算操作，其包含圖形、媒體或計算著色器程式。TD/IC邏輯 223A-223F對子核心內之執行單元執行區域執行緒調度與執行緒控制操作，並促進在子核心之執行單元上執行之執行緒之間的通訊。3D取樣器 225A-225F可讀取紋理或其他3D圖形相關資料到記憶體中。3D取樣器可基於配置的樣本狀態與和與給定紋理相關聯的紋理格式，以不同的方式讀取紋理資料。媒體取樣器 226A-226F可基於與媒體資料相關聯之類型與格式來執行類似的讀取操作。在一實施例中，各圖形子核心 221A-221F可交替地包含統一3D與媒體取樣器。在子核心 221A-221F中每一者內的執行單元上執行之執行緒可使用各子核心內之共享區域記憶體 228A-228F，以使在執行緒群組內執行之執行緒使用晶載記憶體的共用池來進行。

【0067】圖 2C描繪圖形處理單元 (GPU)239，其包含專用的由多個圖形處理資源組成的多個集合，圖形處理資源被設置成多核心群組 240A-240N。雖然僅提供單一個多核心群組 240A的細節，但應瞭解其他的多核心群組 240B-240N可配備有相同或類似的集合之圖形處理資源。

【0068】如圖所示，多核心群組 240A可包含一組圖形核心 243、一組張量核心 244以及一組光線追蹤核心 245。

排程器/調度器 241 排程與調度圖形執行緒以用於在各種核心 243、244 與 245 上執行。當執行圖形執行緒時，一組暫存器堆 242 儲存被核心 243、244、245 所用之運算元值。這些可包含例如，用於儲存整數值之整數暫存器、用於儲存浮點值之浮點暫存器、用於儲存緊縮資料元素(整數及/或浮點資料元素)之向量暫存器、以及用於儲存張量/矩陣值之磚暫存器(tile registers)。在一實施例中，磚暫存器被實現成由向量暫存器組成的多個組合集合。

【0069】一個或多個組合之 Level 1(L1)快取與共享記憶體單元 247 會將圖形資料(諸如，紋理資料、頂點資料、像素資料、光線資料、邊界體積資料等)區域地儲存在各多核心群組 240A 內。一個或多個紋理單元 247 亦可用以執行紋理化操作，諸如紋理映射與取樣。被多核心群組 240A-240N 全部或其一子集所共享之 Level 2(L2)快取 253 儲存圖形資料及/或指令以用於多個並行圖形執行緒。如圖所示，L2 快取 253 可橫跨複數個多核心群組 240A-240N 被共享。一個或多個記憶體控制器 248 將 GPU 239 耦接到記憶體 249，該記憶體可為系統記憶體(例如，DRAM)及/或專用圖形記憶體(例如，GDDR6 記憶體)。

【0070】輸入/輸出(I/O)電路 250 將 GPU 239 耦接到一個或多個 I/O 裝置 252，諸如數位訊號處理器(DSP)、網路控制器或使用者輸入裝置等。晶載互連可用以將 I/O 裝置 252 耦接到 GPU 239 與記憶體 249。I/O 電路 250 之一個或多個 I/O 記憶體管理單元(IOMMU)251 將 I/O 裝置 252 直接耦接

到系統記憶體 249。在一實施例中，IOMMU 251管理多個頁表集，以將虛擬位址映射到系統記憶體 249中的實體位址。在本實施例中，I/O裝置 252、CPU 246與GPU 239可共享相同的虛擬位址空間。

【0071】在一實現中，IOMMU 251支援虛擬化。在此情況下，其可管理第一組頁表，以將訪客/圖形虛擬位址映射到訪客/圖形實體位址，以及管理第二組頁表，以將訪客/圖形實體位址映射到系統/主機實體位址(例如，在系統記憶體 249內)。第一組與第二組頁表集中的每一者的基礎位址可被儲存在控制暫存器中，並在上下文切換上替換掉(例如，以便將相關的一組頁表之存取提供給新的上下文)。雖然在圖 2C中未繪示出，但核心 243、244、245中的每一者及/或多核心群組 240A-240N中的每一者可包含轉譯後備緩衝區(TLB)以快取訪客虛擬到訪客實體轉譯、訪客實體到主機實體轉譯以及訪客虛擬到主機實體轉譯。

【0072】在一實施例中，CPU 246、GPU 239以及I/O裝置 252被整合到單一半導體晶片及/或晶片封裝上。所描繪之記憶體 249可被整合到相同晶片上或可透過晶片外介面耦接到記憶體控制器 248。在一實現中，記憶體 249包含GDDR6記憶體，其與其他實體系統層級記憶體共享相同的虛擬位址空間，雖然本發明之基本原理並不受限於特定的實現。

【0073】在一實施例中，張量核心 244包含經特定設計以執行矩陣運算的複數個執行單元，該矩陣運算為用以

執行深度學習操作之基礎計算操作。舉例而言，同時矩陣乘法運算可用於神經網路訓練與推論。張量核心 244 可透過使用各式各樣運算元精度來執行矩陣處理，該各式各樣運算元精度包含單一精度浮點(例如，32位元)、半精度浮點(例如，16位元)、整數字(16位元)、位元組(8位元)以及半位元組(4位元)。在一實施例中，神經網路實現擷取各顯像的場景之特徵(可能結合來自多個訊框之細節)以建構出一高品質之最終影像。

【0074】在深度學習實現中，平行矩陣乘法工作可能被排程成在張量核心 244 上執行。詳細地，神經網路之訓練需要相當大量之矩陣點積運算。為了處理 $N \times N \times N$ 矩陣乘法之內積公式，張量核心 244 可包含至少 N 個點積處理元件。在矩陣乘法開始之前，一個完整的矩陣被載入到磚暫存器中，以及在 N 個循環中，每個循環載入第二矩陣的至少一行。各循環中有 N 個點積被處理。

【0075】取決於特定的實現，可以不同的精度儲存矩陣元件，包含 16 位元字、8 位元位元組(例如，INT8)和 4 位元半位元組(例如，INT4)。可為張量核心 244 指定不同的精度模式，以確保將最有效的精度用於不同的工作負載(例如，推論工作負載可容忍量化至位元組和半位元組)。

【0076】在一實施例中，光線追蹤核心 245 加速光線追蹤操作，其用於即時光線追蹤實現與非即時光線追蹤實現兩者。詳細地，光線追蹤核心 245 包含光線穿越/相交電路，用於使用邊界體積層次結構(BVH)執行光線遍歷，並

識別光線之間的相交及包含在BVH體積中的基元。光線追蹤核心245亦可包含用於執行深度測試與剔除之電路(例如使用Z緩衝區或類似的配置)。在一種實現中，光線追蹤核心245與本文所述的影像去雜訊技術一同執行遍歷和相交操作，其至少一部分可在張量核心244上執行。舉例而言，在一實施例中，張量核心244實現深度學習神經網路以執行由光線追蹤核心245所產生的訊框之去雜訊。然而，CPU 246、圖形核心243及/或光線追蹤核心245亦可實現全部或部分之去雜訊及/或深度學習演算法。

【0077】此外，如上所述，用以去雜訊之分散式方法可被採用，其中GPU 239是在經由網路或高速互連而耦接至其他計算裝置之計算裝置中。在本實施例中，互連計算裝置共享神經網路學習/訓練資料，以改善整體系統學習對不同類型之影像訊框及/或不同圖形應用程式執行去雜訊之速度。

【0078】在一個實施例中，光線追蹤核心245處理所有的BVH遍歷和光線基元相交，以避免圖形核心243因每條光線有成千上萬條指令而超載。在一實施例中，各光線追蹤核心245包含第一組專用電路，其用於執行邊界框測試(例如，用於遍歷操作)，以及包含第二組專用電路，其用於執行光線三角形相交測試(例如，已遍歷之相交光線)。因此，在一個實施例中，多核心群組240A可簡單地發動光線探測器，並且光線追蹤核心245獨立執行光線遍歷和相交，並將命中資料(例如，命中，無命中，多次命

中等)返回到執行緒上下文。其他的核心 243與 244被釋放以執行其他的圖形或計算工作，而光線追蹤核心 245執行遍歷與相交操作。

【0079】在一實施例中，各光線追蹤核心 245包含遍歷單元以執行BVH測試操作，以及包含執行光線基元相交測試之相交單元。相交單元產生「命中」、「無命中」或「多命中」之回應，其被提供到適當的執行緒。在遍歷與相交操作中，其他核心(例如，圖形核心 243與張量核心 244)之執行資源被釋放，以執行其他形式之圖形工作。

【0080】在下述的特定實施例中，使用一混合光柵化/光線追蹤方法，其中工作被分散在圖形核心 243與光線追蹤核心 245之間。

【0081】在一實施例中，光線追蹤核心 245(及/或其他核心 243與 244)包含對光線追蹤指令集(諸如 Microsoft 的 DirectX 光線追蹤(DXR)，其包含 DispatchRays命令)之硬體支援，以及光線產生、最接近命中、任意命中和未命中著色器，其可針對各物件分配唯一的著色器和紋理之集合。由光線追蹤核心 245、圖形核心 243與張量核心 244所支援之另一光線追蹤平台為 Vulkan 1.1.85。然而，應註明，本發明之基本原理並不受限於任何特定之光線追蹤ISA。

【0082】一般而言，各種核心 245、244、243可支援光線追蹤指令集，其包含用於光線產生、最接近命中、任何命中、光線基元相交、每基元與分層邊界框建構、未中、拜訪與例外之指令/功能。更具體地，一實施例包含

光線追蹤指令，以執行以下功能：

【0083】光線產生-光線產生指令被執行以針對各像素、樣本或其他使用者定義之工作分配。

【0084】最接近命中-最接近命中指令可被執行以定位場景中光線與基元之最近相交點。

【0085】任何命中-任何命中指令識別出在場景內光線與基元之間的多個相交，並可能識別出一個新的最接近相交點。

【0086】相交-相交指令執行光線基元相交測試並輸出結果。

【0087】每基元邊界框建構-此指令在給定基元或基元群組(例如當建構新的BVH或其他加速資料結構時)周圍建構邊界框。

【0088】未中-表示在場景內或在指定區域之場景內，光線未中所有的幾何圖形。

【0089】拜訪-表示光線將遍歷之所有子體積。

【0090】例外-包含各種類型之例外處置器(例如，為各種錯誤條件被調用)。

【0091】圖2D是根據本文所述若干實施例的通用圖形處理單元(GPGPU)270之方塊圖，該GPGPU可配置成圖形處理器及/或計算加速器。GPGPU 270可透過一個或多個系統及/或記憶體匯流排與主機處理器(例如，一個或多個CPU 246)及記憶體271、272互連。在一實施例中，記憶體271係可與一個或多個CPU 246共享的系統記憶體，而記憶

體 272 是專用於 GPGPU 270 的裝置記憶體。在一實施例中，GPGPU 270 與裝置記憶體 272 內之組件可被映射到可由一個或多個 CPU 246 存取之記憶體位址中。對記憶體 271 與 272 之存取可透過記憶體控制器 268 而促進。在一實施例中，記憶體控制器 268 包含內部直接記憶體存取 (DMA) 控制器 269 或可包含用以執行原本由 DMA 控制器執行之操作的邏輯。

【0092】GPGPU 270 包含多個快取記憶體，包含 L2 快取 253、L1 快取 254、指令快取 255 以及共享記憶體 256，其至少一部分可被劃分作為快取記憶體。GPGPU 270 亦包含多個計算單元 260A-260N。各計算單元 260A-260N 包含一組向量暫存器 261、純量暫存器 262、向量邏輯單元 263 以及純量邏輯單元 264。計算單元 260A-260N 亦可包含區域 (local) 共享記憶體 265 以及程式計數器 266。計算單元 260A-260N 可與恆定快取 267 (其可用以儲存恆定資料) 耦接，該恆定資料為在核心運行或在 GPGPU 270 上執行著色器程式期間不會改變的資料。在一實施例中，恆定快取 267 是純量資料快取且快取的資料可直接被提取到純量暫存器 262 中。

【0093】在操作期間，一個或多個 CPU 246 可將命令寫入到已映射到可存取位址空間中的 GPGPU 270 中暫存器或記憶體中。命令處理器 257 可從暫存器或記憶體讀取命令並判定該等命令在 GPGPU 270 內應如何被處理。執行緒調度器 258 可接著用以調度執行緒到計算單元 260A-260N

以執行此等命令。各計算單元260A-260N可獨立於其他的計算單元執行執行緒。額外地，各計算單元260A-260N可被獨立地配置以用於條件計算並可條件地輸出計算之結果到記憶體。當提交命令完成時，命令處理器257可中斷一個或多個CPU 246。

【0094】圖3A-3C描繪由本文所述實施例所提供之附加圖形處理器與計算加速器架構之方塊圖。圖3A-3C之元件，其具有與本文其他任意圖式之元件相同的參考編號(或名稱)者係可以類似於本文其他處所述者的任何方式來運作或運行，但不限於此。

【0095】圖3A是圖形處理器300之方塊圖，其可為離散圖形處理單元或可為整合有複數個處理核心之圖形處理器，或是其他的半導體裝置，諸如但不限於記憶體裝置或網路介面。在若干實施例中，圖形處理器透過記憶體映射之I/O介面來與圖形處理器上之暫存器通訊，並具有設置於處理器記憶體中的命令。在若干實施例中，圖形處理器300包含記憶體介面314以存取記憶體。記憶體介面314可為連接至區域記憶體、一個或多個內部快取、一個或多個共享外部快取及/或系統記憶體之介面。

【0096】在若干實施例中，圖形處理器300亦包含顯示控制器302，以驅動顯示器輸出資料到顯示裝置318。顯示控制器302包含用於一個或多個覆蓋平面之硬體，其用於顯示及組成多層的視訊或使用使用者介面元件。顯示裝置318可為內部或外部顯示裝置。在一實施例中，顯示裝置

318是頭戴式顯示裝置，諸如虛擬實境(VR)顯示裝置或擴增實境(AR)顯示裝置。在若干實施例中，圖形處理器300包含視訊編解碼器引擎306，以將媒體編碼至一個或多個媒體編碼格式、自一個或多個媒體編碼格式解碼或轉碼於一個或多個媒體編碼格式之間，該格式包含但不限於諸如MPEG-2的移動畫面專家群組(MPEG)格式，諸如H.264/MPEG-4 AVC、H.265/HEVC的進階視訊編碼(AVC)格式，開放媒體聯盟(AOMedia)VP8、VP9以及電影電視工程師協會(SMPTE)421M/VC-1和聯合照片專家群組(JPEG)格式(例如JPEG，和運動JPEG(MJPEG)格式)。

【0097】在若干實施例中，圖形處理器300包含區塊影像轉移(BLIT)引擎，以執行二維(2D)光柵器操作，其包含例如位元邊界區塊轉移。然而，在一實施例中，2D圖形操作是使用圖形處理引擎(GPE)310的一個或多個組件來執行。在若干實施例中，GPE 310為用於執行圖形操作(包含三維(3D)圖形操作與媒體操作)之計算引擎。

【0098】在若干實施例中，GPE 310包含用於執行3D操作之3D管線312，該操作諸如使用作用於3D基元形狀(例如，矩形、三角形等等)之處理功能來顯像三維影像與場景。3D管線312包含可程式與固定功能元件，其執行在該元件內之各種任務及/或生成到3D/媒體子系統315之執行緒。雖然在3D管線312可用於執行媒體操作，但GPE 310之實施例亦包含專門地用以執行媒體操作(諸如視訊後處理與影像增強)之媒體管線316。

【0099】在若干實施例中，媒體管線316包含用以執行一個或多個專用媒體操作(諸如，視訊解碼加速、視訊去交錯與視訊編碼加速)之固定功能或可程式邏輯單元，以取代或代表視訊編解碼器引擎306。在若干實施例中，媒體管線316額外地包含執行緒生成單元以生成在3D/媒體子系統315上執行之執行緒。生成的執行緒在3D/媒體子系統315中所包含一個或多個圖形執行單元上執行媒體操作之計算。

【0100】在若干實施例中，3D/媒體子系統315包含用於執行由3D管線312與媒體管線316所生成執行緒之邏輯。在一實施例中，管線向3D/媒體子系統315傳送執行緒執行請求，且3D/媒體子系統315包含用於針對可用執行緒執行資源進行仲裁與調度各種請求之執行緒調度邏輯。執行資源包含圖形執行單元陣列，以處理3D與媒體執行緒。在若干實施例中，3D/媒體子系統315包含用於執行緒指令與資料之一個或多個內部快取。在若干實施例中，子系統亦包含共享記憶體(包含暫存器與可定址記憶體)以在執行緒之間共享資料並用以儲存輸出資料。

【0101】圖3B描繪根據本文所述實施例之具有磚式架構(tiled architecture)的圖形處理器320。在一實施例中，圖形處理器320包含圖形處理引擎群322，其在圖形引擎磚310A-310D中具有多實例之圖3A的圖形處理引擎310。各圖形引擎磚310A-310D可透過一組磚互連323A-323F來互連。各圖形引擎磚310A-310D透過記憶體互連325A-325D

亦可連接至記憶體模組或記憶體裝置 326A-326D。記憶體裝置 326A-326D 可使用任何圖形記憶體技術。舉例而言，記憶體裝置 326A-326D 可為圖形雙重資料率 (GDDR) 記憶體。在一實施例中，記憶體裝置 326A-326D 為高頻寬記憶體 (HBM) 模組，其可與其各別的圖形引擎磚 310A-310D 於晶粒上。在一實施例中，記憶體裝置 326A-326D 為堆疊記憶體裝置，其可堆疊在其各別的圖形引擎磚 310A-310D 的頂部上。在一實施例中，各圖形引擎磚 310A-310D 與相關聯記憶體 326A-326D 位於分離的小晶片上，該小晶片接合到基底晶粒 (base die) 或基底基板 (base substrate)，將進一步於圖 11B-11D 中詳述。

【0102】圖形處理器 320 可被配置使用非一致記憶體存取 (NUMA) 系統，其中記憶體裝置 326A-326D 與相關聯之圖形引擎磚 310A-310D 耦接。給定的記憶體裝置可被圖形引擎磚存取，而非直接與其連接的磚。然而，在存取本地磚 (local tile) 的情況下，對記憶體裝置 326A-326D 的存取延遲可為最低的。在一個實施例中，啟用快取同調 NUMA (ccNUMA) 系統，其使用磚互連 323A-323F 使圖形引擎磚 310A-310D 中的快取控制器之間得以通訊，以在多個快取儲存相同的記憶體位置時保持一致的記憶體影像。

【0103】圖形處理引擎群 322 可與晶載或封裝上構造互連 324 連接。構造互連 324 可賦能在圖形引擎磚 310A-310D 與諸如視訊編解碼器引擎 306 與一個或多個複製引擎 304 等組件之間的通訊。複製引擎 304 可用以將資料移出及

移入記憶體裝置 326A-326D 及圖形處理器 320 外部的記憶體 (例如，系統記憶體)，以及在其之間移動。構造互連 324 亦可用以互連圖形引擎磚 310A-310D。圖形處理器 320 可選地包含顯示控制器 302，以賦能與外部顯示裝置 318 之連接。圖形處理器亦可被配置成圖形或計算加速器。在加速器配置中，顯示控制器 302 與顯示裝置 318 可被省略。

【0104】圖形處理器 320 可透過主機介面 328 來連接至主機系統。主機介面 328 可賦能在圖形處理器 320、系統記憶體及/或其他系統組件之間的通訊。主機介面 328 可例如是 PCI express 匯流排或其他類型的主機系統介面。

【0105】圖 3C 描繪根據本文所述實施例之計算加速器 330。計算加速器 330 在架構上可包含與圖 3B 之圖形處理器 320 相似之處，並且計算加速器被最佳化用於計算之加速。計算引擎群 332 可包含一組計算引擎磚 340A-340D，其包含最佳化用於平行或基於向量之通用計算操作的執行邏輯。在若干實施例中，計算引擎磚 340A-340D 並不包含固定功能圖形處理邏輯，雖然在一實施例中，一個或多個計算引擎磚 340A-340D 可包含用以執行媒體加速之邏輯。計算引擎磚 340A-340D 可透過記憶體互連 325A-325D 以連接至記憶體 326A-326D。記憶體 326A-326D 與記憶體互連 325A-325D 可以是圖形處理器 320 中的類似技術，或可為不同技術。圖形計算引擎磚 340A-340D 亦可透過一組磚互連 323A-323F 而互連並且可透過構造互連 324 來連接及/或互連。在一實施例中，計算加速器 330 包含大的 L3 快取

336，其可被配置成全裝置快取(device-wide cache)。計算加速器330可與圖3B中圖形處理器320類似的方式，透過主機介面328來連接至主機處理器與記憶體。

圖形處理引擎

【0106】圖4係根據若干實施例之圖形處理器之圖形處理引擎410的方塊圖。在一實施例中，圖形處理引擎(GPE)410為圖3A所示GPE 310的一個版本，且亦可表示圖3B之圖形引擎磚310A-310D。圖4之元件，其具有與本文其他任意圖式之元件相同的參考編號(或名稱)者係可以類似於本文其他處所述者的方式來運作或運行，但不限於此。舉例而言，描繪圖3A之3D管線312與媒體管線316。媒體管線316在若干實施例之GPE 410中為可選的，且可能未明確表示包含於GPE 410中。舉例而言，在至少一實施例中，單獨的媒體及/或影像處理器耦接至GPE 410。

【0107】在若干實施例中，GPE 410耦接至或包含命令串流器403，其將命令串流提供至3D管線312及/或媒體管線316。在若干實施例中，命令串流器403係與記憶體(其可為系統記憶體，或一個或多個內部快取記憶體與共享快取記憶體)耦接。在若干實施例中，命令串流器403接收來自記憶體之命令並將該命令傳送至3D管線312及/或媒體管線316。命令係自環形緩衝區直接提取，該環形緩衝區儲存針對3D管線312與媒體管線316之命令。在一實施例中，環形緩衝區可額外地包含批次命令緩衝區(batch

command buffer)，其儲存複數批的多個命令。用於3D管線312之命令亦可包含針對儲存於記憶體中之資料的參照，諸如但不限於針對3D管線312之頂點與幾何資料及/或針對媒體管線316之影像資料與記憶體物件。3D管線312與媒體管線316藉由透過各別管線內的邏輯執行操作或藉由調度一個或多個執行緒至圖形核心陣列414以處理命令與資料。在一實施例中，圖形核心陣列414包含一個或多個區塊之圖形核心(例如，一個或多個圖形核心415A與一個或多個圖形核心415B)，各區塊包含一個或多個圖形核心。各圖形核心包含一組圖形執行資源，其包含用以執行圖形與計算操作之通用與圖形特定執行邏輯，以及固定功能紋理處理及/或機器學習與人工智慧加速邏輯。

【0108】在各種實施例中，3D管線312可包含固定功能與可程式邏輯以藉由處理指令與調度執行緒至圖形核心陣列414，來處理一個或多個著色器程式，諸如頂點著色器、幾何著色器、像素著色器、片段著色器、計算著色器或其他的著色器程式。圖形核心陣列414提供用於處理這些著色器程式的統一執行資源塊。在圖形核心陣列414之圖形核心415A-414B內之多功能執行邏輯(例如，執行單元)包含對各種3D API著色器語言之支援，並可執行與多個著色器相關聯之多個同行執行緒(simultaneous execution thread)。

【0109】在若干實施例中，圖形核心陣列414包含用以執行媒體功能(諸如，視訊及/或影像處理)之執行邏輯。

在一實施例中，執行單元包含可程式之通用邏輯，以除了執行圖形處理操作外亦執行平行通用計算操作。通用邏輯可平行或協同圖 1 之處理器核心 107 或圖 2A 之核心 202A-202N 內之通用邏輯執行處理操作。

【0110】由在圖形核心陣列 414 上執行的執行緒產生之輸出資料可將輸出至統一返回緩衝區(URB)418 中之記憶體。URB 418 可儲存用於多執行緒之資料。在若干實施例中，URB 418 可用以在圖形核心陣列 414 上執行的不同執行緒之間傳送資料。在若干實施例中，URB 418 可額外地被用於在圖形核心陣列上與共享功能邏輯 420 內的固定功能邏輯上的執行緒之間的同步。

【0111】在若干實施例中，圖形核心陣列 414 為可擴縮的，以使得陣列包含可變數量之圖形核心，且基於目標功率與 GPE 410 之性能層級，各具有可變數量之執行單元。在一實施例中，執行資源是動態可擴縮的，以使得執行資源可視需求被啟用或停用。

【0112】圖形核心陣列 414 與包含多個資源之共享功能邏輯 420 耦接，該多個資源被共享於在圖形核心陣列中圖形核心之間。在共享功能邏輯 420 內之共享功能為硬體邏輯單元，其向圖形核心陣列 414 提供專用補充功能。在各種實施例中，共享功能邏輯 420 包含但不限於取樣器 421、數學 422 與執行緒間通訊(ITC)423 邏輯。此外，若干實施例實現在共享功能邏輯 420 內之一個或多個快取 425。

【0113】至少在針對給定之專用功能的需求不足以包

含在圖形核心陣列 414 中的情況下，實現共享功能。反之，專用功能的單一例示被實現為共享功能邏輯 420 中的獨立實體，並在圖形核心陣列 414 中的執行資源之間被共享。在圖形核心陣列 414 之間共享並被包含在圖形核心陣列 414 內之一組精確的功能因實施例而有所不同。在若干實施例中，共享功能邏輯 420 內被圖形核心陣列 414 廣泛使用之特定共享功能可包含於圖形核心陣列 414 內之共享功能邏輯 416 中。在各種實施例中，圖形核心陣列 414 內之共享功能邏輯 416 可包含共享功能邏輯 420 中的若干或全部邏輯。在一實施例中，共享功能邏輯 420 內所有的邏輯元件可於圖形核心陣列 414 之共享功能邏輯 416 中重複。在一實施例中，共享功能邏輯 420 被排除在圖形核心陣列 414 內之共享功能邏輯 416 之外。

執行單元

【0114】圖 5A-5B 描繪執行緒執行邏輯 500，其包含在根據本文所述實施例的圖形處理器核心中採用的處理元件陣列。圖 5A-5B 之元件，其具有與本文其他任意圖式之元件相同的參考編號(或名稱)者係可以類似於本文其他處所述者的方式來運作或運行，但不限於此。圖 5A-5B 描繪執行緒執行邏輯 500 之概述，其可為圖 2B 中各子核心 221A-221F 所描繪之硬體邏輯的表現。圖 5A 為通用圖形處理器內之一個執行單元的表現，而圖 5B 為可用於計算加速器內之一個執行單元的表現。

【0115】如圖5A所描繪，在若干實施例中，執行緒執行邏輯500包含著色器處理器502、執行緒調度器504、指令快取506、可擴縮執行單元陣列(包含複數個執行單元508A-508N)、取樣器510、共享區域記憶體511、資料快取512、以及資料埠514。在一實施例中，藉由基於工作負載之計算需求啟用或停用一個或多個執行單元(例如，執行單元508A、508B、508C、508D、從508N-1到508N中任一者)，可擴縮執行單元陣列可被動態地擴縮。在一實施例中，所包含的組件透過互連構造被互連，該互連構造將各組件連接在一起。在若干實施例中，執行緒執行邏輯500包含經由一個或多個指令快取506、資料埠514、取樣器510與執行單元508A-508N至記憶體(諸如，系統記憶體或快取記憶體)之一個或多個連接。在若干實施例中，各執行單元(例如，508A)為獨立可程式通用計算單元，其能夠在為各執行緒平行處理多個資料元素時，執行多個同時的硬體執行緒。在各種實施例中，執行單元508A-508N之陣列是可擴縮的，以包含任意數量之個別執行單元。

【0116】在若干實施例中，執行單元508A-508N主要用以執行著色器程式。著色器處理器502可處理各種著色器程式，並透過執行緒調度器504，調度與著色器程式相關聯之執行緒。在一實施例中，執行緒調度器包含邏輯，以對來自圖形和媒體管線的執行緒發起請求進行仲裁，並在執行單元508A-508N中的一個或多個執行單元上實例化所請求的執行緒。舉例而言，幾何管線可調度頂點、鑲嵌

(tessellation)或幾何著色器至執行緒執行邏輯以供處理用。在若干實施例中，執行緒調度器 504 亦可處理來自執行著色器程式之運行時間執行緒生成請求。

【0117】在若干實施例中，執行單元 508A-508N 支援包含對許多標準 3D 圖形著色器指令具原生支援之指令集，以使得來自圖形庫(例如，Direct 3D 與 OpenGL)之著色器程式可以最小的轉譯被執行。執行單元支援頂點與幾何處理(例如，頂點程式、幾何程式、頂點著色器)、像素處理(例如，像素著色器、片段著色器)與通用處理(例如，計算與媒體著色器)。各執行單元 508A-508N 能夠執行多發出之單指令多資料(SIMD)執行，並且多執行緒操作能夠在面對更高延遲的記憶體存取時實現高效的執行環境。在各執行單元內之各硬體執行緒具有專用高頻寬暫存器堆與相關聯獨立執行緒狀態。執行係每時脈對管線多發出者，能夠執行整數、單與倍精準度浮點數運算、SIMD 分支能力、邏輯運算、超越操作、以及其他的雜項操作。在等待來自記憶體或共享功能其中之一的資料時，執行單元 508A-508N 內之依賴性邏輯會造成等待執行緒休眠，直到請求資料已經被返回。雖然等待執行緒在休眠，但硬體資源可用於處理其他執行緒。舉例而言，在與頂點著色器操作相關聯之延遲期間，執行單元可執行像素著色器、片段著色器、或其他類型之著色器程式(包含不同頂點著色器)之操作。各種實施例可藉由使用單指令多執行緒(SIMT)來代替 SIMD 的使用或除了 SIMD 的使用以外再使用 SIMT。對 SIMD 核心或

操作之參照亦可應用到SIMT或結合SIMT應用到SIMD。

【0118】在執行單元508A-508N中的各執行單元對多個資料元素陣列進行操作。資料元素之數量是「執行大小」或用於指令之通道的數量。執行通道是執行邏輯單元，用於資料元素存取、遮罩與指令內的流程控制。通道之數量可能獨立於特定圖形處理器之實體算數邏輯單元(ALU)或浮點單元(FPU)之數量。在若干實施例中，執行單元508A-508N支援整數與浮點資料類型。

【0119】執行單元指令集包含SIMD指令。各種資料元素可作為緊縮資料類型被儲存在暫存器中，且執行單元將基於元件之資料大小處理各種元件。舉例而言，當對256位元寬之向量進行操作時，256位元之向量被儲存在暫存器中且執行單元對向量操作如四個分離的54位元緊縮資料元素(四字(Quad-Word，QW)大小資料元素)、八個分離的32位元緊縮資料元素(雙字(Double Word，DW)大小資料元素)、十六個分離的16位元緊縮資料元素(字(Word，W)大小資料元素)或三十二個分離的8位元資料元素(位元組(Byte，B)大小資料元素)。然而，不同的向量寬度與暫存器大小是可能的。

【0120】在一實施例中，一個或多個執行單元可被組合成具有對融合的EU為共用的之執行緒控制邏輯(507A-507N)之融合的執行單元509A-509N，其中執行緒控制邏輯(507A-507N)對融合的EU為共用的。多個EU可被融合至一個EU群組中。融合的EU群組中的各EU可被配置以執行

各自的 SIMD 硬體執行緒。在融合的 EU 群組中 EU 的數量可根據實施例而有所差異。此外，每個 EU 可執行各種 SIMD 寬度，包含但不限於 SIMD8、SIMD16 與 SIMD32。各融合的圖形執行單元 509A-509N 包含至少兩個執行單元。舉例而言，融合的執行單元 509A 包含第一 EU 508A、第二 EU 508B、與對第一 EU 508A 與第二 EU 508B 為共用的之執行緒控制邏輯 507A。執行緒控制邏輯 507A 控制在該融合的圖形執行單元 509A 上執行之執行緒，允許融合的執行單元 509A-509N 內之各 EU 可透過使用共用的指令指標暫存器來進行執行。

【0121】執行緒執行邏輯 500 中包含一個或多個內部指令快取 (例如 506)，以快取用於執行單元的執行緒指令。在一些實施例中，包含一個或多個資料快取 (例如 512) 以在執行緒執行期間快取執行緒資料。在執行邏輯 500 上執行之執行緒亦可明確地儲存管理的資料於共享區域記憶體 511 中。在若干實施例中，包含取樣器 510 以提供用於 3D 操作之紋理取樣以及用於媒體操作之媒體取樣。在若干實施例中，取樣器 510 包含專用紋理或媒體取樣功能，以在取樣處理期間並在提供取樣資料給執行單元之前，處理紋理或媒體資料。

【0122】在執行期間，圖形與媒體管線透過執行緒生成與調度邏輯，傳送執行緒啟動請求到執行緒執行邏輯 500。一旦一群組之幾何物件已經被處理與光柵化成像素資料後，在著色器處理器 502 內之像素處理器邏輯 (例如，

像素著色器邏輯、片段著色器邏輯等)被調用，以進一步計算輸出資訊並使得結果可被寫入到輸出表面(例如，顏色緩衝區、深度緩衝區、模板緩衝區等)。在若干實施例中，像素著色器或片段著色器計算在光柵化物件上內插的各種頂點屬性之值。在若干實施例中，在著色器處理器 502 內之像素處理器邏輯接著執行應用程式介面(API)支援像素或片段著色器程式。為執行著色器程式，著色器處理器 502 透過執行緒調度器 504，調度執行緒到執行單元(例如 508A)。在若干實施例中，著色器處理器 502 使用取樣器 510 中的紋理取樣邏輯，以存取記憶體中所儲存紋理映射圖中的紋理資料。對紋理資料與輸入幾何資料進行之算數運算會計算出針對各幾何片段之像素顏色資料，或從進一步處理中丟棄一個或多個像素。

【0123】在若干實施例中，資料埠 514 對執行緒執行邏輯 500 提供一種記憶體存取機制，以輸出經處理資料到記憶體，用於在圖形處理器輸出管線上之進一步處理。在若干實施例中，資料埠 514 包含或耦接到一個或多個快取記憶體(例如，資料快取 512)，以透過資料埠快取用於記憶體存取之資料。

【0124】在一實施例中，執行邏輯 500 亦可包含光線追蹤器 505，其可提供光線追蹤加速功能。光線追蹤器 505 可支援光線追蹤指令集，其包含用於產生光線之指令/功能。光線追蹤指令集可與由圖 2C 中光線追蹤核心 245 所支援之光線追蹤指令集類似或不同。

【0125】圖 5B 描繪根據實施例的執行單元 508 之例示性內部細節。圖形執行單元 508 可包含指令提取單元 537、通用暫存器堆陣列 (GRF) 524、架構暫存器堆陣列 (ARF) 526、執行緒仲裁器 522、傳送單元 530、分支單元 532、一組 SIMD 浮點單元 (FPU) 534 以及在一實施例中包含一組專用整數 SIMD ALU 535。GRF 524 與 ARF 526 包含與在圖形執行單元 508 中可能活動的 (active) 各同時硬體執行緒相關聯之一組通用暫存器堆與架構暫存器堆。在一實施例中，在 ARF 526 中維持每個執行緒架構狀態，而在執行緒執行期間所用之資料被儲存在 GRF 524 中。各執行緒之執行狀態 (包含各執行緒之指令指標) 可被保存在 ARF 526 中的執行緒特定暫存器中。

【0126】在一實施例中，圖形執行單元 508 具有一種架構，其為同時多重執行緒 (SMT) 與細粒度 (fine-grained) 交錯多重執行緒 (IMT) 之組合。架構具有模組化配置，其可在設計時間基於每執行單元之同時執行緒之目標數量與暫存器之數量來進行微調，其中執行單元資源被劃分予用於執行多個同時執行緒之邏輯。可由圖形執行單元 508 所執行的邏輯執行緒之數量並不受限於硬體執行緒之數量，且多個邏輯執行緒可被分配給各硬體執行緒。

【0127】在一實施例中，圖形執行單元 508 可共同發出多個指令，其可各為不同的指令。圖形執行單元執行緒 508 之執行緒仲裁器 522 可調度指令至傳送單元 530、分支單元 532、或 SIMD FPU 534 其中之一以供執行。各執行緒

可存取 GRF 524內128個通用暫存器，其中各暫存器可儲存32位元組，其可作為32位元資料元素之SIMD 8元件向量被存取。在一實施例中，各執行單元執行緒可存取 GRF 524內之4 Kbytes，雖然實施例未受限於此，且在其他實施例中可提供更多或更少的暫存器資源。在一實施例中，圖形執行單元508被劃分成可獨立執行計算操作之七個硬體執行緒，儘管每執行單元之執行緒數量亦可根據實施例而有所變化。舉例而言，在一實施例中，支援多達16個硬體執行緒。在七個執行緒可存取4Kbytes之實施例中，GRF 524可儲存總共28Kbytes。在16個執行緒可存取4Kbytes之實施例中，GRF 524可儲存總共64Kbytes。彈性定址模式可允許暫存器可一起定址以有效建構較寬的暫存器或用以表示步幅型矩形區塊資料結構(strided rectangular block data structure)。

【0128】在一實施例中，記憶體操作、取樣器操作與其他較長延遲系統通訊透過「傳送」指令而被調度，該傳送指令是被訊息傳遞傳送單元530所執行。在一實施例中，分支指令被調度到專用分支單元532，以促進SIMD發散與最終收斂。

【0129】在一實施例中，圖形執行單元508包含一個或多個SIMD浮點單元(FPU)534以執行浮點運算。在一實施例中，FPU 534亦支援整數計算。在一實施例中，FPU 534可SIMD執行高達M個數量之32位元浮點(或整數)運算，或SIMD執行高達2M個16位元整數或16位元浮點運

算。在一實施例中，FPU之至少其中之一提供擴充數學能力以支援高處理量超越數學函數(high-throughput transcendental math functions)與雙重精度54位元浮點。在若干實施例中，亦存在一組8位元整數SIMD ALU 535，且可專門地被最佳化以執行與機器學習計算相關聯之操作。

【0130】在一實施例中，圖形執行單元508的多個實施例的數個陣列可在圖形子核心群組(例如，子切片)中實例化。為了擴縮性，產品架構師可選擇每子核心群組中執行單元之確切數量。在一實施例中，執行單元508可在複數個執行通道上執行指令。在另一個實施例中，在圖形執行單元508上執行之各執行緒係在不同通道上執行。

【0131】圖6描繪根據實施例的附加執行單元600。執行單元600可為計算最佳化執行單元，例如用於圖3C之計算引擎磚340A-340D中，但不限於此。執行單元600之變體亦可用於如圖3B中之圖形引擎磚310A-310D中。在一實施例中，執行單元600包含執行緒控制單元601、執行緒狀態單元602、指令提取/預取單元603、以及指令解碼單元604。執行單元600額外地包含暫存器堆606，其儲存可被分配給執行單元內之數個硬體執行緒的數個暫存器。執行單元600額外地包含傳送單元607與分支單元608。在一實施例中，傳送單元607與分支單元608之操作可類似於圖5B中圖形執行單元508之傳送單元530與分支單元532。

【0132】執行單元600亦包含計算單元610，其包含多個不同類型之功能單元。在一實施例中，計算單元610包

含 ALU 單元 611，其包含算數邏輯單元陣列。ALU 單元 611 可被配置以執行 64 位元、32 位元與 16 位元之整數與浮點運算。整數與浮點運算可同時被執行。計算單元 610 亦可包含脈動陣列 (systolic array) 612 與數學單元 613。脈動陣列 612 包含寬度 W 與深度 D 之資料處理單元網路，其可用於以脈動方式執行向量或其他資料平行操作。在一實施例中，脈動陣列 612 可被配置以執行矩陣運算，諸如矩陣點積運算。在一實施例中，脈動陣列 612 支援 16 位元浮點運算，亦支援 8 位元與 4 位元整數運算。在一實施例中，脈動陣列 612 可被配置以加速機器學習操作。在這樣的實施例中，脈動陣列 612 可被配置以支援浮點 16 位元浮點格式。在一實施例中，可包含數學單元 613 以利用比 ALU 單元 611 更有效與更低功率方式執行數學運算的特定子集。數學單元 613 可包含數學邏輯之變體，其可在由其他實施例提供的圖形處理引擎的共享功能邏輯中找到 (例如，圖 4 的共享功能邏輯 420 之數學邏輯 422)。在一實施例中，數學單元 613 可被配置以執行 32 位元與 64 位元之浮點運算。

【0133】執行緒控制單元 601 包含用以控制執行單元內之執行緒之執行的邏輯。執行緒控制單元 601 可包含執行緒仲裁邏輯，以將執行單元 600 內執行緒之執行開始、停止與搶占。執行緒狀態單元 602 可用以儲存針對分配以在執行單元 600 上執行的執行緒之執行緒狀態。將執行緒狀態儲存在執行單元 600 內可在該等執行緒阻塞或閒置時以便執行緒之快速搶占。指令提取/預取單元 603 可從較高

層級執行邏輯之指令快取(如圖 5A 之指令快取 506)提取指令。指令提取/預取單元 603 亦可基於對正在執行執行緒之分析，發出預取請求以將指令載入到指令快取中。指令解碼單元 604 可用以解碼將由計算單元執行之指令。在一實施例中，指令解碼單元 604 可用作次級解碼器，以將複雜指令解碼為組成微操作(constituent micro-operation)。

【0134】執行單元 600 額外地包含暫存器堆 606，其可被在執行單元 600 上執行之硬體執行緒使用。暫存器堆 606 中的暫存器被劃分予用以執行在執行單元 600 之計算單元 610 內之多個同時執行緒之邏輯。可由圖形執行單元 600 所執行的邏輯執行緒之數量並不受限於硬體執行緒之數量，且多個邏輯執行緒可被分配給各硬體執行緒。基於支援硬體執行緒之數量，暫存器堆 606 的大小在不同實施例可有所不同。在一實施例中，暫存器重新命名可用以將暫存器動態分配給硬體執行緒。

【0135】圖 7 是描繪根據若干實施例之圖形處理器指令格式 700 之方塊圖。在一個或多個實施例中，圖形處理器執行單元支援具有多種格式之指令的指令集。實線框描繪通常被包含在執行單元指令中的組件，而虛線則包含可選或僅包含在指令之子集合中的組件。在若干實施例中，所描述且描繪之指令格式 700 為巨集指令，因為其為供應至執行單元之指令而非在指令被處理後由指令解碼產生之微操作。

【0136】在若干實施例中，圖形處理器執行單元原生

支援在 128 位元指令格式 710 中的指令。基於所選指令、指令選項與運算元數量，64 位元壓縮指令格式 730 可用於若干指令。原生 128 位元指令格式 710 提供對所有指令選項之存取，而若干選項與操作在 64 位元格式 730 中有所限制。因實施例的不同，64 位元格式 730 中可用的原生指令可能有所變化。在若干實施例中，透過使用索引欄位 713 中一組索引值，指令部分被壓縮。執行單元硬體基於索引值參照一組壓縮表並使用壓縮表輸出以重建原生指令成 128 位元指令格式 710。可使用其他的指令大小與格式。

【0137】針對各種格式，命令運算碼(opcode) 712 定義執行單元將執行之操作。執行單元在各運算元之多個資料元素上平行執行各指令。舉例而言，因應加法指令，執行單元在表示紋理元件或畫面元件之各顏色通道上執行同時加法運算。在默認情況下，執行單元在運算元之所有資料通道執行各指令。在若干實施例中，指令控制欄位 714 賦能對特定執行選項之控制，諸如通道選擇(例如，預測)與資料通道順序(例如，調配(swizzle))。對於 128 位元指令格式 710 的指令，執行大小欄位(exec-size field) 716 限制了將平行執行之資料通道的數量。在若干實施例中，執行大小欄位 716 不可用於 64 位元壓縮指令格式 730。

【0138】若干種執行單元指令可具有多達三個運算元，包含兩個源運算元 src0 720 與 src1 722 及一個目的地運算元 718。在若干實施例中，執行單元支援雙目的地指令，其中該等目的地其中之一為隱含的。資料調處指令可

具有第三個源運算元(例如，SRC2 724)，其中命令運算碼 712判定該源運算元之數量。指令的最後一個源運算元可為立即(例如，硬編碼)值，其與指令一起傳遞。

【0139】在若干實施例中，128位元指令格式710包含存取/位址模式欄位726，其指定例如是使用直接暫存器定址模式或使用間接暫存器定址模式。當使用直接暫存器定址模式時，藉由指令中的位元直接提供一個或多個運算元之暫存器位址。

【0140】在若干實施例中，128位元指令格式710包含存取/位址模式欄位726，其指定指令之位址模式及/或存取模式。在一實施例中，存取模式用以定義指令之資料存取對準。若干實施例支援包含16位元組對準存取模式與1位元組對準存取模式等存取模式，其中存取模式之位元組對齊判定指令運算元之存取對齊。舉例而言，當在第一模式中時，針對源運算元與目的地運算元，指令可使用位元組對準之定址，而在第二模式中時，針對所有的源運算元與目的地運算元，指令可使用16位元組對準之定址。

【0141】在一實施例中，存取/位址模式欄位726之位址模式部分(address mode portion)判定該指令是使用直接或間接定址。當使用直接暫存器定址模式時，指令中的位元直接提供一個或多個運算元之暫存器位址。當使用間接暫存器定址模式時，可基於指令中位址暫存器值與位址立即欄位來計算出一個或多個運算元之暫存器位址。

【0142】在若干實施例中，基於運算碼712位元欄位

來將指令分組，以簡化運算碼解碼 740。針對 8 位元運算碼，第 4、5、與 6 位元允許執行單元可判定運算碼類型。所示之精準運算碼分組僅是一範例。在若干實施例中，移動與邏輯運算碼群組 742 包含資料移動與邏輯指令(例如，移動(mov)與比較(cmp))。在若干實施例中，移動與邏輯群組 742 共享五個最高有效位元(MSB)，其中移動(mov)指令的形式是 0000xxxxb 以及邏輯指令的形式是 0001xxxxb。流控制(flow control)指令群組 744(例如，呼叫，跳越(jmp))包含形式為 0010xxxxb(例如，0x20)的指令。雜項指令群組 746 包含指令之混合，包含形式為 0011xxxxb(例如，0x30)的同步指令(例如，等待、傳送)。平行數學指令群組 748 包含形式為 0100xxxxb(例如，0x40)的分量式(component-wise)算數指令(例如，加法、乘法(mul))。平行數學群組 748 在資料通道上平行執行算數運算。向量數學群組 750 包含形式為 0101xxxxb(例如，0x50)的算數指令(例如，dp4)。向量數學群組對向量運算元執行諸如點積計算等算數。在一實施例中，所描繪的運算碼解碼 740 可用以判定執行單元中何部分將用以執行解碼的指令。舉例而言，若干指令可被指定為脈動指令，其將由脈動陣列所執行。其他指令(諸如，光線追蹤指令(未繪示出))可被路由至在執行邏輯的一切片或一分區中之光線追蹤核心或光線追蹤邏輯。

圖形管線

【0143】圖8係另一實施例的圖形處理器800之方塊圖。圖8之元件，其具有與本文其他任意圖式之元件相同的參考編號(或名稱)者係可以類似於本文其他處所述者的方式來運作或運行，但不限於此。

【0144】在若干實施例中，圖形處理器800包含幾何管線820、媒體管線830、顯示引擎840、執行緒執行邏輯850以及顯像輸出管線870。在若干實施例中，圖形處理器800為多核心處理系統中的圖形處理器，該系統包含一個或多個通用處理核心。圖形處理器係藉由暫存器寫入到一個或多個控制暫存器(未繪示出)或透過環形互連802發送至圖形處理器800之命令所控制。在若干實施例中，環形互連802將圖形處理器800耦接到其他的處理組件，諸如其其他的圖形處理器或通用處理器。來自環形互連802之命令被命令串流器803解譯，以及命令串流器803提供指令至幾何管線820或媒體管線830中的個別組件。

【0145】在若干實施例中，命令串流器803將頂點提取器805之操作導向，該頂點提取器805從記憶體讀取頂點資料並執行由命令串流器803所提供的頂點處理命令。在若干實施例中，頂點提取器805提供頂點資料至頂點著色器807，其對各頂點執行座標空間轉換與打光操作。在若干實施例中，藉由透過執行緒調度器831來調度執行緒至執行單元852A-852B，頂點提取器805與頂點著色器807執行頂點處理指令。

【0146】在若干實施例中，執行單元852A-852B係向

量處理器陣列，其具有用於執行圖形與媒體操作之指令集。在若干實施例中，執行單元 852A-852B 具有附接的 L1 快取 851，其對各陣列為特定的或在多個陣列之間為共享的。快取可被配置成資料快取、指令快取或單一快取，其被劃分以在不同分區中包含資料與指令。

【0147】在若干實施例中，幾何管線 820 包含鑲嵌組件以執行 3D 物件之硬體加速鑲嵌。在若干實施例中，可程式外殼著色器 811 配置鑲嵌操作。可程式域著色器 817 提供鑲嵌輸出之後端評估。鑲嵌器 (tessellator) 813 在外殼著色器 811 之方向上操作並包含特殊用途邏輯，以基於粗糙幾何模型產生一組詳細幾何物件，該粗糙幾何模型被作為輸入提供到幾何管線 820。在若干實施例中，若無使用鑲嵌，則鑲嵌組件 (例如，外殼著色器 811、鑲嵌器 813 與域著色器 817) 可被略過。

【0148】在若干實施例中，完整的幾何物件可透過被調度到執行單元 852A-852B 之一個或多個執行緒被幾何著色器 819 處理，或者可直接進入至剖切器 (clipper) 829。在若干實施例中，幾何著色器對整個幾何物件進行操作，而非如在先前階段之圖形管線般僅對數個頂點或數個頂點斑塊進行操作。若鑲嵌被停用，則幾何著色器 819 接收來自頂點著色器 807 之輸入。在若干實施例中，幾何著色器 819 為可由幾何著色器程式程式化，以在鑲嵌單元被停用之情況下執行幾何鑲嵌。

【0149】在光柵化之前，剖切器 829 處理頂點資料。

剖切器 829 可為固定功能剖切器或具有剖切與幾何著色器功能之可程式剖切器。在若干實施例中，在顯像輸出管線 870 中的光柵器與深度測試組件 873 調度像素著色器以將幾何物件轉換成每個像素表示。在若干實施例中，像素著色器邏輯被包含在執行緒執行邏輯 850 中。在若干實施例中，應用程式可略過光柵器與深度測試組件 873，並透過串流輸出單元 823 存取未經光柵化的頂點資料。

【0150】圖形處理器 800 具有互連匯流排、互連構造或若干其他互連機制，其允許資料與訊息在處理器之主要組件之間傳遞。在若干實施例中，透過資料埠 856，執行單元 852A-852B 及相關聯的邏輯單元(例如，L1 快取 851、取樣器 854、紋理快取 858 等)互連，以執行記憶體存取以及與處理器之顯像輸出管線組件之通訊。在若干實施例中，取樣器 854、快取 851、858 與執行單元 852A-852B 每一者具有各自的記憶體存取路徑。在一實施例中，紋理快取 858 亦可被配置為取樣器快取。

【0151】在若干實施例中，顯像輸出管線 870 包含光柵器與深度測試組件 873，其將基於頂點物件轉換成相關聯之基於像素表示。在若干實施例中，光柵器邏輯包含窗口器/遮罩器單元以執行固定功能之三角形與線光柵化。在若干實施例中亦可使用相關聯之顯像快取 878 與深度快取 879。像素操作組件 877 對資料執行基於像素之操作，雖然在某些情況下，與 2D 操作相關聯之像素操作(例如，具混和 (blending) 之位元區塊影像轉移)是由 2D 引擎 841 所執

行，或在顯示時間上由顯示控制器 843 使用覆蓋顯示平面予以替代。在若干實施例中，所有圖形組件可使用共享 L3 快取 875，允許資料共享而無需使用主系統記憶體。

【0152】在若干實施例中，圖形處理器媒體管線 830 包含媒體引擎 837 與視訊前端 834。在若干實施例中，視訊前端 834 從命令串流器 803 接收管線命令。在若干實施例中，媒體管線 830 包含單獨的命令串流器。在若干實施例中，視訊前端 834 在將命令傳送至媒體引擎 837 之前先處理媒體命令。在若干實施例中，媒體引擎 837 包含執行緒生成功能，用以生成用以透過執行緒調度器 831 被調度到執行緒執行邏輯 850 之執行緒。

【0153】在若干實施例中，圖形處理器 800 包含顯示引擎 840。在若干實施例中，顯示引擎 840 係在處理器 800 外部並透過環形互連 802 或若干其他互連匯流排或構造而耦接到圖形處理器。在若干實施例中，顯示引擎 840 包含 2D 引擎 841 與顯示控制器 843。在若干實施例中，顯示引擎 840 包含特殊用途邏輯，其能夠獨立於 3D 管線操作。在若干實施例中，顯示控制器 843 與顯示裝置(未繪示出)耦接，該顯示裝置可為系統整合顯示裝置，諸如筆記型電腦中者，或是透過顯示裝置連接器予以附接之外部顯示裝置。

【0154】在若干實施例中，幾何管線 820 與媒體管線 830 可被配置以基於多個圖形與媒體程式介面來執行操作，並且沒有特定於任一種應用程式介面(API)。在若干

實施例中，用於圖形處理器之驅動器軟體將API呼叫(其特定於特定圖形或媒體庫)轉譯成可被圖形處理器處理之命令。在若干實施例中，針對開放圖形庫(OpenGL)、開放計算語言(OpenCL)及/或Vulkan圖形和計算API所提供之支援全部來自於Khronos集團。在若干實施例中，針對Direct3D庫所提供之支援亦來自於微軟公司。在若干實施例中，可支援此等庫之組合。針對開源電腦視覺庫(OpenCV)亦提供支援。如果可從未來的API的管線進行到圖形處理器的管線之映射，則更將支援具有相容3D管線之未來的API。

圖形管線程式設計

【0155】圖9A係描繪根據若干實施例之圖形處理器命令格式900之方塊圖。圖9B係描繪根據一實施例之圖形處理器命令序列910之方塊圖。圖9A中實線框描繪通常被包含在圖形命令中的組件，而虛線則包含可選或僅包含在圖形命令之子集合中的組件。圖9A之例示性圖形處理器指令格式900包含用以識別客戶端902、命令運算碼(opcode)904與用於命令之資料906的資料欄位。子運算碼(sub-opcode)905與命令大小908亦可被包含在若干命令中。

【0156】在若干實施例中，客戶端902指定處理命令資料的圖形裝置之客戶端單元。在若干實施例中，圖形處理器命令剖析器檢查各命令之客戶端欄位，以調節命令之進一步處理並路由該命令資料至適當的客戶端單元。在若

于實施例中，圖形處理器客戶端單元包含記憶體介面單元、顯像單元、2D單元、3D單元以及媒體單元。各客戶端單元具有對應的處理管線，其處理該等命令。一旦命令被客戶端單元接收，客戶端單元讀取運算碼 904 及子運算碼 905 (若存在)，以判定要執行之操作。客戶端單元使用資料欄位 906 中的資訊以執行命令。針對若干命令，明確 (explicit) 的命令大小 908 被預期用以指定命令之大小。在若干實施例中，基於命令運算碼，命令剖析器自動判定命令中至少若干者之大小。在若干實施例中，透過多個雙重字來將命令對準。可使用其他的命令格式。

【0157】圖 9B 中的流程圖描繪例示性圖形處理器命令序列 910。在若干實施例中，具有圖形處理器的一實施例的資料處理系統之軟體或韌體使用所示之命令序列的一版本，以設定、執行以及終結一組圖形操作。僅為了例示目的繪示出並描述作為樣本的一命令序列，因為實施例並不限於此等特定命令或此命令序列。再者，該等命令可以命令序列中的一批命令發出，以便圖形處理器將至少部分地同時處理該命令序列。

【0158】在若干實施例中，圖形處理器命令序列 910 可以管線刷新命令 912 開始，以使得任何活動的圖形管線完成用於該管線的當前未完成的命令。在若干實施例中，3D 管線 922 與媒體管線 924 不會同時地運作。執行管線刷新以使活動的圖形管線完成任何未完成的命令。因應管線刷新，用於圖形處理器之命令的剖析器將暫停命令處理，直

到活動的繪圖引擎完成未完成的操作且相關讀取快取失效。可選地，顯像快取中任何被標記為「骯髒(dirty)」之資料可被刷新到記憶體。在若干實施例中，管線刷新命令 912 可用於管線同步或在將圖形處理器置於低電力狀態之前使用。

【0159】在一些實施例中，當命令序列要求圖形處理器顯式在管線之間切換時，使用管線選擇命令 913。在若干實施例中，在發出管線命令之前，在執行上下文中僅需要一次管線選擇命令 913，除非該上下文是對雙管線發送命令。若干實施例中，在透過管線選擇命令 913 進行管線切換之前，立即需要管線刷新命令 912。

【0160】在若干實施例中，管線控制命令 914 配置用於操作之圖形管線，並且用以程式化 3D 管線 922 及媒體管線 924。在若干實施例中，管線控制命令 914 針對活動的管線，配置管線狀態。在一實施例中，管線控制命令 914 係用於管線同步且用以在處理一批指令之前從活動的管線內的一個或多個快取記憶體中清除資料。

【0161】在若干實施例中，返回緩衝區狀態命令 916 用以針對各別的管線，配置一組返回緩衝區以寫入資料。若干管線操作需要分配、選擇、或配置一個或多個返回緩衝區，以在處理期間寫入中間資料至返回緩衝區中。在若干實施例中，圖形處理器亦使用一個或多個返回緩衝區以儲存輸出資料並執行跨執行緒通訊。在若干實施例中，返回緩衝區狀態 916 包含選擇要使用於一組管線操作的返回

緩衝區之大小與數量。

【0162】基於用於操作之活動的管線，在命令序列中剩餘命令而有所不同。基於管線判定 920，將命令序列定制成開頭為 3D 管線狀態 930 的 3D 管線 922，或開頭為媒體管線狀態 940 的媒體管線 924。

【0163】用以配置 3D 管線狀態 930 之命令包含 3D 狀態設定命令，其係針對頂點緩衝區狀態、頂點元件狀態、恆定顏色狀態、深度緩衝區狀態以及其他在處理 3D 基元命令之前需要配置之狀態變數。至少部分基於所用之特定的 3D API，來判定此等命令之值。在若干實施例中，3D 管線狀態 930 的命令亦能夠選擇性停用或略過特定管線元件，若這些這些元件將不會被使用。

【0164】在若干實施例中，3D 基元 932 的命令係用以提交 3D 基元，以由 3D 管線處理。透過 3D 基元 932 的命令傳遞到圖形處理器之命令與相關聯參數轉送到圖形管線中的頂點提取函數(vertex fetch function)。頂點提取函數使用 3D 基元 932 的命令資料來產生頂點資料結構。頂點資料結構儲存於一個或多個返回緩衝區中。在若干實施例中，透過頂點著色器，3D 基元 932 的命令用以對 3D 基元執行頂點操作。為了處理頂點著色器，3D 管線 922 調度著色器執行緒到圖形處理器執行單元。

【0165】在若干實施例中，透過執行 934 的命令或事件而觸發 3D 管線 922。在若干實施例中，暫存器寫入觸發命令執行。在若干實施例中，透過命令序列中的「go」或

「kick」指令，來觸發執行。在一實施例中，透過使用管線同步命令來觸發指令執行，以經由圖形管線刷新命令序列。3D管線會對3D基元執行幾何處理。一旦完成操作，則產生的幾何物件被光柵化且像素引擎將產生的像素上色。針對此等操作，亦可包含用以控制像素著色與像素後端操作之額外命令。

【0166】在若干實施例中，當執行媒體操作時，圖形處理器命令序列910遵循媒體管線924路徑。一般而言，針對媒體管線924之特定用途與程式設計之方式係取決於要執行的媒體或計算操作。在媒體解碼期間，特定媒體解碼操作可被卸載到媒體管線。在若干實施例中，亦可透過略過媒體管線，且藉由使用由一個或多個通用處理核心所提供之資源來整體或部分地執行媒體解碼。在一實施例中，媒體管線亦包含用於通用圖形處理器單元(GPGPU)操作之元件，其中圖形處理器用以使用計算著色器程式來執行SIMD向量操作，該計算著色器程式係非明確地關於圖形基元顯像。

【0167】在若干實施例中，以與3D管線922類似的方式配置媒體管線924。用以配置媒體管線狀態940之一組命令被調度或置於命令佇列中在媒體物件命令942之前。在若干實施例中，用於媒體管線狀態940之命令包含配置用以處理媒體物件之媒體管線元件的資料。此包含配置在組態媒體管線內之視訊解碼與視訊編碼邏輯的資料，諸如編碼或解碼格式。在若干實施例中，用於媒體管線狀態940

之命令亦支援使用一個或多個指向「間接」狀態元件之指標，其中「間接」狀態元件包含一批狀態設定。

【0168】在若干實施例中，媒體物件命令 942 提供指向媒體物件之指標，以便由媒體管線進行處理。媒體物件包含記憶體緩衝區，其包含待處理之視訊資料。在一些實施例中，在發布媒體物件命令 942 之前，所有媒體管線狀態必須為有效的。一旦管線狀態被配置且媒體物件命令 942 被佇列後，透過執行命令 944 或等效的執行事件（例如，暫存器寫入）來觸發媒體管線 924。來自媒體管線 924 之輸出可接著由 3D 管線 922 或媒體管線 924 所提供之操作進行後處理。在若干實施例中，以與媒體操作類似的方式，配置且執行 GPGPU 操作。

圖形軟體架構

【0169】圖 10 描繪根據若干實施例針對資料處理系統 1000 之例示性圖形軟體架構。在若干實施例中，軟體架構包含 3D 圖形應用程式 1010、作業系統 1020、以及至少一處理器 1030。在若干實施例中，處理器 1030 包含圖形處理器 1032 與一個或多個通用處理器核心 1034。圖形應用程式 1010 與作業系統 1020 各自在資料處理系統之系統記憶體 1050 中執行。

【0170】在若干實施例中，3D 圖形應用程式 1010 包含一個或多個著色器程式，該著色器程式包含著色器指令 1012。著色器語言指令可為高階著色器語言，諸如

Direct3D、OpenGL 著色器語言 (GLSL) 等高階著色器語言 (HLSL)。應用程式亦包含適用於通用處理器核心 1034 執行的機器語言中之可執行指令 1014。應用程式亦包含由頂點資料定義的圖形物件 1016。

【0171】在若干實施例中，作業系統 1020 是微軟公司之 Microsoft® Windows® 作業系統、專屬類 UNIX 作業系統或使用 Linux 核心 (kernel) 之變體的開源類 UNIX 作業系統。作業系統 1020 可支援圖形 API 1022，諸如：Direct3D API、OpenGL API、或 Vulkan API。當使用 Direct3D API 時，作業系統 1020 使用前端著色器編譯器 1024 以將任何 HLSL 之著色器指令 1012 編譯成較低階著色器語言。該編譯可為即時 (JIT) 編譯或應用程式可執行著色器預編譯。在若干實施例中，高階著色器可在 3D 圖形應用程式 1010 之編譯期間被編譯成低階著色器。若干實施例中，著色器指令 1012 係以中間形式被提供，諸如由 Vulkan API 使用之標準可攜式中間表示法 (SPIR) 的一個版本。

【0172】在若干實施例中，使用者模式圖形驅動器 1026 包含後端著色器編譯器 1027，以將著色器指令 1012 轉換成硬體特定表示法。當使用 OpenGL API 時，GLSL 高階語言之著色器指令 1012 被傳遞到使用者模式圖形驅動器 1026 以用於編譯。在若干實施例中，使用者模式圖形驅動器 1026 使用作業系統核心模式功能 1028 來與核心模式圖形驅動器 1029 通訊。在若干實施例中，核心模式圖形驅動器 1029 與圖形處理器 1032 進行通訊，以調度命令和指令。

IP核心實現

【0173】至少一實施例之一個或多個態樣可藉由儲存在機器可讀取媒體上之代表性程式碼來實現，其表示及/或定義諸如處理器等積體電路之中的邏輯。舉例而言，機器可讀取媒體可包含表示處理器內各種邏輯之指令。當被機器讀取時，指令可使該機器製造邏輯以執行本文所述之技術。此等表示法(稱為「IP核心」)係是針對積體電路之可重複使用邏輯單元，其可被儲存在有形機器可讀取媒體上作為描述該積體電路結構之硬體模型。硬體模型可被供應到各種客戶或生產設備，其載入該硬體模型至生產該積體電路之製造機器上。積體電路可被製造以令該電路執行與本文所述任意實施例所描述之相關聯的操作。

【0174】圖 11A 係描繪 IP 核心開發系統 1100 之方塊圖，該系統可用以生產積體電路以執行根據實施例之操作。IP 核心開發系統 1100 可用以產生模組化、可重複使用設計，該設計可被併入至更大的設計中或用以建構整個積體電路(例如，SoC 積體電路)。設計設備 1130 可產生在高階程式設計語言(例如，C/C++)之 IP 核心設計之軟體模擬 1110。透過使用模擬模型 1112，軟體模擬 1110 可用以設計、測試並驗證該 IP 核心之行為。模擬模型 1112 可包含功能、行為及/或計時模擬。接著，可自模擬模型 1112 建立或合成暫存器轉移層級(RTL)設計 1115。RTL 設計 1115 為積體電路之行為的抽象，它模擬硬體暫存器之間的數位訊

號之流動，包含使用模擬的數位訊號所執行之相關聯的邏輯。除了RTL設計1115以外，亦可建立、設計或合成在邏輯層級或電晶體層級處之較低階設計。因此，初始設計與模擬在特定細節上可能會有所不同。

【0175】RTL設計1115或等效物可進一步被設計設備合成為硬體模型1120，其可使用硬體描述語言(HDL)或若干種其他實體設計資料的表示法。可進一步模擬或測試HDL，以驗證IP核心設計。透過使用非揮發性記憶體1140(例如，硬碟、快閃記憶體或任何非揮發性儲存媒體)，IP核心設計可被儲存以供交付至第三方協力廠商製造設備1165。替代地，IP核心設計可經由有線連接1150或無線連接1160傳送(例如，透過網際網路)。製造設備1165可接著至少部分基於IP核心設計製造積體電路。製造的積體電路可被配置以執行根據本文所述至少一實施例之操作。

【0176】圖11B描繪根據本文所述若干實施例之積體電路封裝總成1170之橫截面側視圖。積體電路封裝總成1170描繪本文所述一個或多個處理器或加速器裝置之實現。封裝總成1170包含連接至基板1180之硬體邏輯1172、1174的多個單元。邏輯1172、1174可至少部分於可配置邏輯或固定功能邏輯硬體中實現，並且可包含處理器核心、圖形處理器或其他本文所述加速器裝置任意者的一個或多個部分。邏輯1172、1174之各單元可在半導體晶粒中實現並透過互連結構1173與基板1180耦接。互連結構1173可被

配置以在邏輯 1172、1174與基板 1180之間路由電訊號，並且可包含諸如但不限於凸塊或柱狀等互連。在若干實施例中，互連結構 1173可被配置以路由電訊號，諸如例如輸入/輸出(I/O)訊號及/或與邏輯 1172、1174操作相關聯的電力或接地訊號。在若干實施例中，基板 1180為基於環氧樹脂的層板基板。在其他實施例中，基板 1180可包含其他適當類型之基板。透過封裝互連 1183，封裝總成 1170可連接至其他電子裝置。封裝互連 1183可耦接到基板 1180之表面，以將電訊號路由到其他的電子裝置，諸如主機板、其他晶片組或多晶片模組。

【0177】在若干實施例中，邏輯 1172、1174之多個單元電性耦接到橋接器 1182，其被配置以在邏輯 1172、1174之間路由電訊號。橋接器 1182可為密集的互連結構，其為提供對電訊號之路由。橋接器 1182可包含橋接器基板，其由玻璃或適當的半導體材料構成。可在橋接器基板上形成電性路由特徵，以在邏輯 1172與 1174之間提供晶片對晶片之連接。

【0178】雖然僅描繪兩個單元之邏輯 1172、1174與橋接器 1182，但本文所述實施例可在一個或多個晶粒上包含更多或更少邏輯單元。該一個或多個晶粒可透過零個或更多橋接器連接，如在邏輯被包含在單一晶粒上之情況下，橋接器 1182可被排除。替代地，多個晶粒或邏輯之單元可透過一個或多個橋接器連接。額外地，多個邏輯單元、晶粒與橋接器可以其他可能的配置連接在一起，包含三維配

置。

【0179】圖 11C 描繪封裝總成 1190，其包含連接至基板 1180(例如，基底晶粒)之硬體邏輯小晶片之多個單元。本文所述之圖形處理單元、平行處理器、及/或計算加速器可由各自生產之多樣化的矽小晶片所構成。在此上下文中，小晶片為一種至少部分封裝之積體電路，其包含不同的邏輯單元且可與其他小晶片組裝成為更大的封裝。具有不同 IP 核心邏輯之多樣化的一組小晶片可組裝成一個單一裝置。額外地，透過使用主動式中介層技術，小晶片可整合至基底晶粒中或基底小晶片中。在此所述之概念使 GPU 內不同形式的 IP 之間能夠互連與通訊。IP 核心可使用不同處理技術來生產且可在生產期間中進行組合，這避免集中多個 IP 至相同生產處理中的複雜度(特別是在具有多種特點 IP 之大 SoC 上)。支持使用多種處理技術可改善上市所需的時間，並提供一種具有成本效益的方式來建立多個產品 SKU。額外地，分開的 IP 更適合獨立地進行電源關斷(可關閉在給定工作負載中未使用的組件)而降低整體功耗。

【0180】硬體邏輯小晶片可包含特殊用途硬體邏輯小晶片 1172、邏輯或 I/O 小晶片 1174、及/或記憶體小晶片 1175。硬體邏輯小晶片 1172 與邏輯或 I/O 小晶片 1174 可至少部分於可配置邏輯或固定功能邏輯硬體中實現，並且可包含處理器核心、圖形處理器、平行處理器、或其他本文所述加速器裝置任意者的一個或多個部分。記憶體小晶片 1175 可為 DRAM(例如，GDDR、HBM)記憶體或快取

(SRAM)記憶體。

【0181】各小晶片可被製造為單獨的半導體晶粒，並且透過互連結構 1173 與基板 1180 耦接。互連結構 1173 可被配置以在各種小晶片與基板 1180 內邏輯之間路由電訊號。互連結構 1173 可包含諸如但不限於凸塊或柱狀等互連。在若干實施例中，互連結構 1173 可被配置以路由電訊號，諸如例如輸入/輸出(I/O)訊號及/或與邏輯、I/O與記憶體小晶片操作相關聯的電力或接地訊號。

【0182】在若干實施例中，基板 1180 為基於環氧樹脂的層板基板。在其他實施例中，基板 1180 可包含其他適當類型之基板。透過封裝互連 1183，封裝總成 1190 可連接至其他電子裝置。封裝互連 1183 可耦接到基板 1180 之表面，以將電訊號路由到其他的電子裝置，諸如主機板、其他晶片組或多晶片模組。

【0183】在若干實施例中，邏輯或 I/O 小晶片 1174 與記憶體小晶片 1175 可透過橋接器 1187 電性耦接，該橋接器被配置以在邏輯或 I/O 小晶片 1174 與記憶體小晶片 1175 之間路由電訊號。橋接器 1187 可為密集的互連結構，其為提供對電訊號之路由。橋接器 1187 可包含橋接器基板，其由玻璃或適當的半導體材料構成。可在橋接器基板上形成電性路由特徵，以在邏輯或 I/O 小晶片 1174 與記憶體小晶片 1175 之間提供晶片對晶片之連接。橋接器 1187 亦可被稱為矽橋接器或互連橋接器。舉例而言，在若干實施例中，橋接器 1187 為嵌入式多晶粒互連橋接器(EMIB)。在若干實施

例中，橋接器 1187 可簡單地為從一小晶片至另一小晶片的直接連接。

【0184】基板 1180 可包含用於 I/O 1191、快取記憶體 1192 與其他硬體邏輯 1193 之硬體組件。構造 1185 可嵌入至基板 1180 內以使基板 1180 內各種邏輯小晶片與邏輯 1191、1193 之間通訊。在一實施例中，I/O 1191、構造 1185、快取、橋接器與其他硬體邏輯 1193 可整合至基底晶粒中，該基底晶粒置於基板 1180 的頂部。構造 1185 可為晶片上網路 (network on a chip) 的互連或另一種形式的分封交換構造，其於封裝總成的組件之間交換資料封包。

【0185】在各種實施例中，封裝總成 1190 可包含藉由構造 1185 或一個或多個橋接器 1187 互連之較少或較多數量之組件與小晶片。封裝總成 1190 內之小晶片可以 3D 或 2.5D 的配置而設置。一般而言，橋接器結構 1187 可用以促進在例如邏輯或 I/O 小晶片與記憶體小晶片之間點對點的互連。構造 1185 可用以將各種邏輯及/或 I/O 小晶片 (例如，小晶片 1172、1174、1191、1193) 與其他邏輯及/或 I/O 小晶片進行互連。在一實施例中，基板內之快取記憶體 1192 可作為用於封裝總成 1190 之全局快取、作為分散式全局快取的一部分或作為用於構造 1185 之專用快取。

【0186】圖 11D 描繪根據實施例之包含可互換小晶片 1195 的封裝總成 1194。可互換小晶片 1195 可被組裝至一個或多個基底小晶片 1196、1198 上的標準化插槽中。基底小晶片 1196、1198 可透過橋接器互連 1197 耦接，該橋接器可

與本文所述其他橋接器互連類似且可例如是 **EMIB**。記憶體小晶片亦可透過橋接器互連而連接至邏輯或 **I/O** 小晶片。**I/O**與邏輯小晶片可透過互連構造通訊。基底小晶片可各支援一個或多個標準化格式之插槽，以用於邏輯或 **I/O**或記憶體/快取其中之一。

【0187】在一實施例中，**SRAM**與供電電路可被製造成一個或多個基底小晶片 1196、1198，其可使用不同於堆疊在基底小晶片頂部上之可互換小晶片 1195 的處理技術來製造。舉例而言，基底小晶片 1196、1198 可使用較大的製程技術製造，而可互換小晶片可使用較小的製程技術生產。一個或多個可互換小晶片 1195 可為記憶體(例如，**DRAM**)小晶片。基於針對使用封裝總成 1194 之產品的電力及/或性能，可為封裝總成 1194 選擇不同的記憶體密度。額外地，基於針對產品之電力及/或性能，可在組裝時選擇具有類型數量不同的功能單元之邏輯小晶片。額外地，包含不同類型 **IP** 邏輯核心之小晶片可插入至可互換小晶片插槽中，以使混合處理器(hybrid processor)設計能夠混合匹配(mix and match)不同技術的 **IP** 區塊。

示例性系統單晶片積體電路

【0188】圖 12-13 描繪根據本文所述各種實施例之示例性積體電路與相關聯的圖形處理器，其可使用一個或多個 **IP** 核心製造。除了所描繪示的之外，可包含其他邏輯與電路(包含附加圖形處理器/核心、周邊介面控制器或通用

處理器核心)。

【0189】圖12係根據實施例的描繪例示性系統單晶片積體電路1200之方塊圖，其可使用一個或多個IP核心製造。例示性積體電路1200包含一個或多個應用處理器1205(例如，CPU)、至少一圖形處理器1210以及可額外地包含影像處理器1215及/或視訊處理器1220，上述任一者可為來自相同或多個不同設計設備之模組化IP核心。積體電路1200包含周邊或匯流排邏輯，其包含USB控制器1225、UART控制器1230、SPI/SDIO控制器1235以及I²S/I²C控制器1240。額外地，積體電路可包含與高解析多媒體介面(HDMI)控制器1250與行動產業處理器介面(MIPI)顯示介面1255其中之一或多者耦接之顯示裝置1245。可透過包含快閃記憶體與快閃記憶體控制器的快閃記憶體子系統1260來提供儲存。透過記憶體控制器1265可提供記憶體介面，以存取SDRAM或SRAM記憶體裝置進行。若干積體電路額外地包含嵌入式安全引擎1270。

【0190】圖13A-13B係根據本文所述實施例描繪用於SoC內之例示性圖形處理器之方塊圖。圖13A係根據實施例描繪系統單晶片積體電路之例示性圖形處理器1310，其可使用一個或多個IP核心製造。圖13B係根據實施例描繪系統單晶片積體電路之附加的例示性圖形處理器1340，其可使用一個或多個IP核心製造。圖13A之圖形處理器1310為低功率圖形處理器核心之一範例。圖13B之圖形處理器1340為較高性能圖形處理器核心之一範例。圖形處理器

1310、1340每一者可為圖12之圖形處理器1210的變體。

【0191】如圖13A示出，圖形處理器1310包含頂點處理器1305與一個或多個片段處理器1315A-1315N(例如，1315A、1315B、1315C、1315D到1315N-1及1315N)。圖形處理器1310可透過單獨的邏輯執行不同著色器程式，以使得頂點處理器1305被最佳化以執行針對頂點著色器程式之操作，而一個或多個片段處理器1315A-1315N執行為片段或像素著色器程式執行片段(例如，像素)著色操作。頂點處理器1305執行3D圖形管線之頂點處理階段，並產生基元與頂點資料。片段處理器1315A-1315N使用由頂點處理器1305所產生之基元與頂點資料，以產生在顯示裝置上顯示之訊框緩衝區。在一實施例中，片段處理器1315A-1315N被最佳化以執行如在OpenGL API所提供之片段著色器程式，其可用以執行類似於Direct 3D API中提供的像素著色器程式之操作。

【0192】圖形處理器1310額外地包含一個或多個記憶體管理單元(MMU)1320A-1320B、快取1325A-1325B、及電路互連1330A-1330B。一個或多個MMU 1320A-1320B為圖形處理器1310(包含對頂點處理器1305及/或片段處理器1315A-1315N)提供虛擬對實體位址映射，其除了在一個或多個快取1325A-1325B中儲存的頂點或影像/紋理資料以外，可參照記憶體中儲存之頂點或影像/紋理資料。在一實施例中，一個或多個MMU 1320A-1320B可與系統內其他MMU同步，包含與圖12之一個或多個應用處理器1205、

影像處理器 1215 及 / 或 視訊處理器 1220 相關聯的一個或多個 MMU，以使各處理器 1205-1220 可參與共享或統一虛擬記憶體系統。根據實施例，一個或多個電路互連 1330A-1330B 使圖形處理器 1310 能夠與 SoC 內其他 IP 核心介接 (不論是透過 SoC 之內部匯流排或透過直接連接)。

【0193】如圖 13B 所示，圖形處理器 1340 包含圖 13A 中圖形處理器 1310 之一個或多個記憶體管理單元 (MMU) 1320A-1320B、快取 1325A-1325B、及電路互連 1330A-1330B。圖形處理器 1340 包含一個或多個著色器核心 1355A-1355N (例如，1355A、1355B、1355C、1355D、1355E、1355F 至 1355N-1 及 1355N)，其提供統一著色器核心架構，其中單一核心或類型或核心可執行所有類型之可程式著色器碼，包含著色器程式碼以實現頂點著色器、片段著色器及 / 或計算著色器。著色器核心的確切數量在不同實施例與實現上可有所不同。額外地，圖形處理器 1340 包含核心間任務管理器 1345，其作為執行緒調度器以將執行緒調度到一個或多個著色器核心 1355A-1355N 與拼貼單元 1358，以加速用於基於磚顯像之拼貼操作，其中針對一場景的顯像操作在影像空間中細分，例如，以利用場景中區域空間同調性或以最佳化內部快取之使用。

機器學習工作負載之張量加速邏輯

【0194】圖 14 係根據實施例的資料處理系統 1400 之方塊圖。資料處理系統 1400 係一種具有處理器 1402、統一記

記憶體 1410 與包含機器學習加速邏輯的 GPGPU 1420 之異質處理系統。處理器 1402 與 GPGPU 1420 可為本文所述任意種處理器與 GPGPU/平行處理器。處理器 1402 可執行儲存於系統記憶體 1412 中用於編譯器 1415 的指令。編譯器 1415 在處理器 1402 上執行，以將原始碼 1414A 編譯成編譯碼 1414B。編譯碼 1414B 可包含可由處理器 1402 執行之指令及/或可由 GPGPU 1420 執行之指令。在編譯期間，編譯器 1415 可執行操作以插入元資料，包含存在編譯碼 1414B 中的資料平行化層級的提示及/或與關於與基於編譯碼 1414B 調度的執行緒相關聯的資料局部性的提示。編譯器 1415 可包含執行這些操作所必須的資訊，或這些操作可在運行時間庫 1416 的輔助下執行。運行時間庫 1416 亦可輔助編譯器 1415 編譯原始碼 1414A，亦可包含在運行時間中與編譯碼 1414B 連接的指令，以促進在 GPGPU 1420 上執行編譯指令 (compiled instruction)。

【0195】統一記憶體 1410 表示一個統一位址空間，其可被處理器 1402 與 GPGPU 1420 存取。統一記憶體可包含系統記憶體 1412 以及 GPGPU 記憶體 1418。GPGPU 記憶體 1418 為在 GPGPU 1420 位址空間內之記憶體，並可包含系統記憶體 1412 之若干部分或全部。在一實施例中，GPGPU 記憶體 1418 亦可包含至少一部分之任何由 GPGPU 1420 專用之記憶體。在一實施例中，儲存在系統記憶體 1412 中的編譯碼 1414B 可被映射到 GPGPU 記憶體 1418 中，以供 GPGPU 1420 存取。

【0196】 GPGPU 1420包含多個計算區塊1424A-1424N，其可包含本文所述的各種處理資源其中之一或多者。處理資源可為或包含各種不同的計算資源，例如執行單元、計算單元、串流多處理器、圖形多處理器或多核心組。在一實施例中，GPGPU 1420額外地包含張量(例如：矩陣)加速器1423，其可包含一個或多個特殊功能計算單元，此係被設計以加速多個矩陣運算(例如，點積等)中的一子集。張量加速器1423亦可被稱為張量加速器或張量核心。在一實施例中，張量加速器1423內的邏輯組件可分佈於多個計算區塊1424A-1424N的處理資源中。

【0197】 GPGPU 1420亦可包含一組可被計算區塊1424A-1424N與張量加速器1423共享的資源，其包含但不限於一組暫存器1425、電源與效能模組1426與快取1427。在一實施例中，暫存器1425包含直接與間接可存取之暫存器，其中，間接可存取暫存器被最佳化以由張量加速器1423使用。電源與效能模組1426可被配置以為計算區塊1424A-1424N調整對計算區塊1424A-1424N內之電閘閒置組件的供電與時脈頻率。在各種實施例中，快取1427可包含指令快取及/或較低層級資料快取。

【0198】 GPGPU 1420可額外地包含L3資料快取1430，其可用以快取被張量加速器1423及/或計算區塊1424A-1424N內計算元件自統一記憶體1410存取之資料。在一實施例中，L3資料快取1430包含可由計算區塊1424A-1424N內計算元件與張量加速器1423共享之共享區域記憶

體 1432。

【0199】在一實施例中，GPGPU 1420包含指令處理邏輯，諸如提取與解碼單元(fetch and decode unit)1421以及排程器控制器(scheduler controller)1422。提取與解碼單元1421包含提取單元與解碼單元，用於提取與解碼指令以供計算區塊1424A-1424N其中之一或多者或張量加速器1423執行。透過排程器控制器1422，指令可被排程到計算區塊1424A-1424N內適當的功能單元或張量加速器。在一實施例中，排程器控制器1422為ASIC，其可配置以執行進階排程操作。在一實施例中，排程器控制器1422為微控制器或每指令低功耗處理核心(low energy-per-instruction processing core)，其能夠執行從韌體模組載入之排程器指令。

【0200】在一實施例中，可以直接將計算區塊1424A-1424N執行的一些功能排程到或卸載至張量加速器1423。在各種實施例中，張量加速器1423包含處理元件邏輯，該處理元件邏輯被配置以有效地執行矩陣計算運算，諸如由3D圖形或計算著色器程式所用之乘法與加法運算與點積運算等。在一實施例中，張量加速器1423可被配置以加速由機器學習框架使用之操作。在一實施例中，張量加速器1423為一種明確地被配置以執行特定一組平行矩陣乘法及/或加法運算之特殊應用積體電路。在一實施例中，張量加速器1423為一種提供固定功能邏輯之現場可程式閘極陣列(FPGA)，該固定運作邏輯可在工作負載之間更新。相對

於計算區塊 1424A-1424N 所能執行的操作，張量加速器 1423 所能執行的該組矩陣運算可能較為侷限。然而，相對於計算區塊 1424A-1424N，張量加速器 1423 能以顯著較高的處理量執行這些操作。

【0201】圖 15 描繪根據一實施例，由指令管線 1500 所執行之矩陣運算 1505。在一實施例中，指令管線 1500 可被配置以執行矩陣運算 1505，諸如，但不限於，點積運算。兩個向量之點積是一個純量值，其等於此兩個向量對應分量的乘積之總和。點積可如下公式(1)所示計算出。

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^n a_i b_i = a_1 b_1 + \dots + a_n b_n \quad (1)$$

【0202】點積可用於卷積神經網路(CNN)的卷積運算。圖 15 描繪二維(2D)卷積，其使用包含點積運算的矩陣運算 1505。雖然描繪 2D 卷積，但可使用 N 維度過濾器對 N 維體積執行 N 維度之卷積。接受域磚(receptive field tile) 1502 強調輸入體積緩衝區(input volume buffer) 1504 中輸入體積的一部分。輸入體積緩衝區可被儲存於記憶體 1530 中。點積矩陣運算 1505 可在接受域磚 1502 中的資料與卷積過濾器中的資料之間執行，以在輸出緩衝區 1506 中產生資料點，該輸出緩衝區亦可被儲存在記憶體 1530 中。記憶體 1530 可為本文所述任何記憶體，包含如圖 14 中所述系統記憶體 1412、GPGPU 記憶體 1418、或一個或多個快取記憶體 1427 與 1430。

【0203】輸出緩衝區 1506 中的資料點之組合表示由卷

積操作所產生之激活映射。激活映射中的每個點都是透過在輸入體積緩衝區 1504 上滑動接受域磚而產生。激活映射資料可被輸入至激活函數以判定輸出激活值。在一實施例中，在框架中，輸入體積緩衝區 1504 之卷積可定義為高階矩陣運算 1505。該高階矩陣運算可透過基元運算(諸如基礎線性代數子程式(BLAS)運算)而執行。基元運算可透過指令管線 1500 執行的硬體指令來加速。

【0204】用以加速硬體指令之指令管線 1500 可包含指令提取與解碼單元 1421 與排程器控制器 1422，該指令提取與解碼單元 1421 可提取與解碼硬體指令而該排程器控制器 1422 可排程已解碼的指令至計算區塊 1424A-1424N 內的一個或多個處理資源及 / 或張量加速器 1423。在一實施例中，可將硬體指令排程至計算區塊 1424A-1424N 及卸載到張量加速器 1423。執行矩陣運算 1505 的一個或多個硬體指令與相關聯的資料可被儲存於記憶體 1530 中。硬體指令的輸出亦可被儲存於記憶體 1530 中。

【0205】在一實施例中，張量加速器 1423 可執行一個或多個硬體指令以使用整合的脈動陣列 1508 (DP 邏輯) 來執行矩陣運算 1505。脈動陣列 1508 可包含可程式與固定功能硬體的組合，其可被配置以執行點積運算。雖然計算區塊 1424A-1424N 內之功能單元亦可被配置以執行點積運算，但脈動陣列 1508 可被配置以相對於計算區塊 1424A-1424N 具有顯著較高的處理量來執行侷限的點積運算子集合。

使用具反饋輸入之脈動陣列之擴縮稀疏矩陣乘法加速

【0206】本文描述使用具反饋輸入之脈動陣列實現擴縮稀疏矩陣乘法加速之架構。該架構在資料顯示大量零之工作負載中加速脈動矩陣乘法。這種架構易於擴縮，保留為了利用自工作負載的資料中發現的稀疏度而採用之最佳化所受到的利益，並容許增加指令吞吐量。相對於其他脈動陣列，優點包含面積縮小、功耗降低以及性能提升。藉由減少在高度稀疏的資料負載下的計算數量，此架構改善脈動點積累加運算的性能。與先前的架構相比，計算量減少使得功耗降低。藉由簡化脈動陣列的設計，該架構也比現有的脈動陣列有更好的擴縮。簡化的設計使該架構可廣泛地併入至加速器設計中，以增加那些加速器的矩陣處理吞吐量。

【0207】用於機器學習演算法之計算的資料顯示大量作為輸入元素的零。在神經網路應用中，此現象係由實現的網路之拓撲及建模的神經元之特性自然地造成。舉例而言，在全互連神經網路中，神經元的行或層的輸出連接至下一層中各神經元的輸入。網路由許多層組成，各層可能都具有大量的神經元。計算神經元之輸出的最後階段係執行激活函數。當神經元之輸入的計算為正值時，此函數通常輸出對正函數(positive function)的求值(evaluation)，而當神經元之輸入的計算為負值時，輸出為零。因此，神經元的大量輸出結果可為零，並且在神經網路之計算的下一步，饋送至下一個網路層。

【0208】在執行具有大量層和神經元的神經網路時，每一層由神經元之輸入的權重值之矩陣和神經元之輸入值之矩陣來表示。神經元的所有輸入都乘以其權重，然後加到神經元之其他輸入的結果中。然後，套用激活函數以得到饋送至下一個網路層的神經元的輸出。為了快速計算神經元的輸入和權重的乘法和加法，許多種技術已於硬體中使用。與本發明有關的係使用脈動陣列進行乘法和累加運算。

【0209】圖 16 描繪包含乘法器與加法器電路的脈動陣列 1600，其以管線的形式組織。第一輸入矩陣的輸入 1612A-1612H 由標記為 Src1 和 Src1+1 到 Src1+7 之輸入中所包含的資料元素表示。輸入 1610A-1610B 對應至第二個輸入矩陣，並標記為 Src2。輸入 1602 (其可能包含初始累加器值) 可作為 Src0 提供。脈動陣列 1600 的處理元件 1611A-1611H 作為一管線結構運作，並且每個級 (stage) 在個時脈週期執行。在每個週期中，每個級都可接收一個新的 src2 輸入，其可與一個新的 Src1 輸入或一個較舊的 Src1 輸入進行計算。一個 Src2 輸入與八個 Src1 輸入一起運算 (例如，每級一個 Src1 輸入)。Src2 輸入之通道 (channel) 的資料元素在處理元件 1611A-1611H 的所有通道廣播。然後，處理元件與 Src1 輸入的所有通道一起運算 Src2 通道。在第一個時脈週期中，Src1 輸入與 Src2 的第一通道的資料元素一起運算。在下一個週期中，第二個 Src1 (標記為 Src1+1) 與 Src2 的第二個通道的資料元素一起運算。此序列在管線的 8 個

級上重複。每個級都將其運算加到前一級的輸出中。在整個管線級中，多個 Src2 輸入以管線的形式進行運算。隨著第一 Src2 輸入的連續通道通過管線級，可在第一級提供新的 Src2 輸入。

【0210】最後一級的輸出 1622 標記為 Dst。其中，d 為脈動深度，e 為每個通道的資料元素數量，通道的輸出由下面的公式(2)描述：

$$Dst_i = Src0_i + \sum_{j=0}^d \sum_{k=0}^e (Src1 + j)_{\text{element } k \text{ of channel } i} * Src2_{\text{element } k \text{ of channel } j} \quad (2)$$

【0211】如公式(2)所示，每個通道可包含多個資料元素，對其平行執行運算。在一個實施例中，儘管可為每個通道配置不同數量的元素，但是每個通道表示四個元素資料向量。在一個實施例中，通道內的資料元素的數量可基於每個資料元素的大小而有所變化。舉例而言，可使用每個元素具有 8 位元資料類型的四個元素向量，兩個具有 16 位元資料類型的元素向量，八個具有 4 位元資料類型(例如 INT4)的元素向量或 16 個具有 2 位元資料類型(例如 INT2)的元素向量來執行點積。可根據 Src1 和 Src2 的資料類型自動調整通道數。指令更可指定欲用於該指令之所需的脈動深度。

【0212】在一個實施例中，處理元件 1611A-1611H 可直接自通用暫存器堆讀取輸入 1610A-1610H、1612A-1612H。在一個實施例中，脈動陣列 1600 包含邏輯，用以自通用暫存器堆讀取輸入 1610A-1610H、1612A-1612H，

並將輸入資料儲存在脈動陣列內部的暫存器、緩衝區或記憶體中。然後，內部邏輯可將輸入資料元素饋送到處理元件 1611A-1611H 進行處理。可將輸出 1622 寫入到脈動陣列 1600 的內部暫存器或記憶體中及/或直接寫入通用暫存器堆中。

【0213】在一個實施例中，當判定輸入至一個級中的乘法器/加法器之元素為零時，可略過乘法/加法而僅傳播先前的輸入。當輸入矩陣是稀疏的(例如，包含大量的零)時，需要執行的運算量就會減少。一些實現可略過乘法/累加級，以避免消耗功率進行乘法/加法運算。然而，簡單地略過運算會導致脈動鏈中存在閒置級。舉例而言，如果 Src2 輸入的元素全為零，則略過與該元素相關聯的該級會節省電源。然而，在該週期中將不執行任何運算。因此，運算的吞吐量將仍然未提高。其他實現可能會重新排列輸入，如此一來，可略過陣列的整個部分。由於添加額外的硬體來重新排列輸入，因此重新排列輸入會增加擴充成本。架構必須考慮輸入缺乏稀疏度的情況，其將會需要使用管線的全部深度。因此，應該將架構設計為所有級隨時都可供使用。

【0214】本文在各種實施例中描述最佳化的架構，以在避免上述問題的同時以某種方式處理稀疏輸入。實施例提供了一種模組化脈動陣列，其可容易地擴縮以滿足不同產品的需求，並允許僅計算非零元素，而無需額外的硬體或閒置時脈週期。這些概念將在此說明書接下來的部分循

序漸進地描述。

具反饋輸入之矩陣乘法加速器

【0215】脈動陣列 1600 為具累加之八個深度乘法/加法陣列，且可稱為 DPAS(點積累加脈動)陣列。這種架構的深度有益於某些產品、專案及/或使用案例。其他產品、專案及/或使用案例可能不需要這樣深度的陣列，且將因陣列具有較低吞吐量、較小面積及/或較低功耗而從中受益。另外，其他產品、專案及/或使用案例可能會受益於藉由使用較深的管線深度而獲得的較高的 DPAS 吞吐量。為了減少脈動陣列中使用的硬體數量，同時保留用於更深層陣列的相同程式設計模型，將反饋輸入加入至深度減小版本的脈動陣列中如圖 16 所示。

【0216】圖 17A-17B 描繪使用四個深度的脈動陣列 1700 來計算八個脈動級(systolic stage)的等效陣列。圖 17A 顯示從外部源接收 Src0 輸入並使用 Src1 和 Src2 輸入處理前四個級的陣列。陣列的輸出被反饋至圖 17B 所示的第二步驟。圖 17B 顯示使用包含已處理值以及 Src1 和 Src2 輸入的反饋資料來計算接下來的四個級。

【0217】如圖 17A 所示，脈動陣列 1700 可接收輸入 1602(如 Src0 輸入)，並透過資料選擇器 1704 讀取(1702)輸入 1602。資料選擇器 1704 在輸入 1602 和反饋輸入 1706 之間進行選擇。處理元件 1611A-1611D 可以類似於脈動陣列 1600 的方式處理輸入 1610A-1610D 和 1612A-1612D。如果

四個級足以完成運算，則處理元件 1611D 可透過資料選擇器 1724 將輸出 1622 寫入 (1722) 至指定的 Dst 暫存器或記憶體。在需要更多級的情況下，資料選擇器 1724 可寫入反饋輸出 1726，其作為反饋輸入 1706 提供至處理元件 1611A。

【0218】如圖 17B 所示，在一個實施例中，反饋輸入 1706 可由處理元件 1611A-1611D 進一步處理。反饋輸入 1706 包含已處理值。在一個實施例中，反饋輸入 1706 更可包含輸入 1610E-1610H、輸入 1612E-1612H，其可在處理前四個級時預取 (pre-fetch)。資料選擇器 1704 選擇反饋輸入 1706 以作為處理元件 1611A 的輸入。接著，處理元件 1611A-1611D 可處理輸入 1610E-1610H 及 1612E-1612H。隨後，資料選擇器 1724 可將第八級結果作為輸出 1622 寫入 (1722) 至指定的 Dst 暫存器。

【0219】圖 16 和圖 17A-17B 的陣列的時序圖如圖 18A-18B 所示。

【0220】圖 18A-18B 顯示兩個時序圖 1800、1850。圖 18A 顯示與圖 16 所示的 8 個深度的脈動陣列 1600 相對應的時序圖 1800。圖 18B 顯示與圖 17A-17B 所示的具反饋輸入之脈動陣列相對應的時序圖 1850。多個時脈週期如圖所示。

【0221】如圖 18A 所示，時序圖 1800 顯示讀取週期 1802，其對應於脈動陣列 1600 的讀取邏輯讀取欲被饋送至陣列的輸入。脈動週期 1804 與如何在陣列的每個級中處理這些輸入相對應。寫入週期 1806 對應至輸出邏輯，其將輸出寫入至指定的目的地位置。在讀取週期 1802 的週期 0 中

讀取之輸入由脈動陣列在脈動週期 1804 的週期 0-7 中處理。在讀取週期 1802 的週期 1 中讀取之輸入，在脈動週期 1804 的週期 1-8 中處理。在脈動週期 1804 的週期 0 開始之處理在寫入週期 1806 的週期「N」中輸出。在一個實施例中，N 的值可與脈動陣列的深度相關。在脈動週期 1804 的週期 1 中開始之處理在週期 1-8 中進行處理並在寫入週期 1806 的週期 N+1 中輸出。由於陣列的管線性質，在週期 N 和 N+1 產生的輸出之計算將由脈動陣列的各個管線級(級 1 到級 8)平行執行。

【0222】讀取週期 1802 的週期領先脈動週期 1804 一個或多個週期。舉例而言，讀取週期 1802 的週期 0 可出現在脈動週期 1804 的週期 0 之前。在一個實施例中，讀取週期 1802 的週期 1 可與脈動週期 1804 的週期 0 同時地出現。在脈動週期 1804 的週期 0 中，脈動陣列的級 1 基於在讀取週期的週期 0 中讀取的輸入來計算 $\text{Src0} + \text{Src1} * \text{Src2.0}$ 。Src1 的元素 0-7 中的每一元素都與 Src2 的元素 0 平行相乘，並加到對應的 Src0 的元素 0-7 中。第 1 級產生的結果將傳遞到第 2 級。在脈動週期 1804 的週期 1 中，級 2 累加級 1 在週期 0 中計算的結果，其結果為 $[\text{Src1} + 1] * \text{Src2.1}$ ，其中 Src2 的元素 1 乘以 $[\text{Src1} + 1]$ 的每個元素。Src1 保留在級 1，這樣在週期 1，級 1 計算具下一個 Src0 輸入及 Src2 輸入之 $[\text{Src0} + 1] + \text{Src1} * [\text{Src2} + 1]$ 。在級 1 中，Src1 的元素 0-7 中的每個元素與 $[\text{Src2} + 1]$ 的元素 0 平行相乘，並加到對應的 $[\text{Src0} + 1]$ 的元素中。對於每個級和每個週期，處理皆以這種模式繼續進行，並

自寫入週期1806的週期N開始，從級8輸出結果。

【0223】如圖18B所示，時序圖1850顯示圖17A-17B的脈動陣列1700的處理過程，自脈動週期1854的週期0開始對在讀取週期1852的週期0讀取的第一組輸入進行處理，並以與圖16中脈動陣列1600的八個深度管線的前四個級相同的方式進行。可理解的是，在圖18B中，讀取週期1852不一定呈現與脈動週期1854中對應的週期對準。脈動陣列1700的讀取邏輯可以與脈動陣列1600相同的方式讀取第一組輸入。在讀取週期1852的週期0至3中讀取的前四個輸入可藉由脈動陣列1700以管線的形式處理。在寫入週期1856中標記為「N」的週期中，產生輸出。因此，在脈動陣列1700之週期N的第一輸出的延遲與在脈動陣列1600相同。然而，由於反饋造成在寫入週期1856的週期N+8至N+11中寫入之Dst3和Dst4的輸出之間存在延遲，使得脈動陣列1700的吞吐量降低。

【0224】對於脈動陣列1700，反饋始於脈動週期1854的週期4。反饋開始直到週期7。一旦反饋始於脈動週期1854的週期4，處理元件只讀取Src1輸入，如脈動週期1854的週期4、5、6和7之虛線輸入所表示。下一組src0和src2輸入將由處理元件從脈動週期1854的週期8開始讀取。讀取邏輯可延遲對Src0和Src2輸入的讀取，直到需要這些輸入為止，或者如圖所示，可連同Src1輸入一起讀取這些輸入。一旦讀取輸入，這些輸入可以由脈動陣列緩衝並重新使用。

【0225】相對於脈動陣列 1600，具反饋之矩陣乘法加速器(脈動陣列 1700)的優點可總結如下：脈動陣列 1700 可使用較少的硬體來計算與脈動陣列 1600類似的管線深度。脈動陣列 1700允許使用與脈動陣列 1600相同的指令，因而允許為脈動陣列 1600開發的工作負載在脈動陣列 1700中重複使用。脈動陣列 1700因使用更少的管線級，功耗更小。相對於脈動陣列 1600，針對讀取和寫入，脈動陣列 1700可在減小的頻寬下運作。脈動陣列 1700可以任何數量的級來實現，儘管使用二的倍數可能是最佳的，以為了凸顯以下所示的實施例的特徵。雖然整個脈動陣列 1700的架構有所減少，但可藉由平行實現脈動陣列 1700的多個實例來達成與脈動陣列 1600相同的吞吐量。

具反饋輸入之可擴縮的矩陣乘法加速器

【0226】第二實施例藉由使用平行單元使用同時指令執行來增加的吞吐量。乘法加速器的多個實例或路徑平行運行。這些實例可共享 Src1，或可具有獨立的 Src1輸入。每個路徑會有自己的 Src2輸入及 Src0輸入。這些實例會具有自己的 Src2輸入及 Src0輸入。圖 19顯示了深度為四個級的二個路徑的版本。另外，圖 20顯示使用深度為二個級的四個路徑的版本。

【0227】圖 19描繪二路徑矩陣乘法加速器 1900，其中每一者路徑的深度為 4 個級。該二路徑矩陣乘法加速器 1900包含用於 Src0輸入的輸入邏輯 1902A-1902B，用於儲

存從輸入邏輯 1910A-1910B接收的資料元素的輸入緩衝區 1911A-1911B，以及用於 Src1以儲存從共享輸入邏輯 1912接收的資料元素的輸入緩衝區 1913A-1913B。每個級包含一對可平行操作的處理元件。第一級包含處理元件 1931A-1931B，第二級包含處理元件 1932A-1932B，第三級包含處理元件 1933A-1933B，第四級包含處理元件 1934A-1934B。每個處理元件 1931A-1931B、1932A-1932B、1931A-1933B、1934A-1934B的硬體邏輯可與脈動陣列 1600或脈動陣列 1700的處理元件(例如，處理元件 1611A-1611D)的硬體邏輯相同或類似，並且可以相同的製程技術或更先進的製程技術生產。相對於脈動陣列 1600的實現，該二路徑矩陣乘法加速器 1900的處理元件更可以更高的頻率運作。處理元件可使用更先進的製程技術來生產。

【0228】可使用與資料選擇器 1704、1724相同或類似的資料選擇器來實現反饋。取決於讀取邏輯的配置，輸入資料可預先被預取至輸入緩衝區中，或者在輸入至處理元件 1931A-1931B之前的一個或多個週期從二路徑矩陣乘法加速器 1900內的暫存器或快取中讀取。第四級的處理元件 1934A-1934B可反饋至對應的第一級的處理元件 1931A-1931B中。動態邏輯深度可以四的倍數實現。在配置邏輯級的數量後，結果可由輸出邏輯 1922A-1922B寫入至指定的目的地。

【0229】圖 20描繪四路徑矩陣乘法加速器 2000，其中每一者路徑的深度為 2 級。四路徑矩陣乘法加速器 2000包

含與二路徑矩陣乘法加速器 1900 相同數量的處理元件，而處理元件被配置以具有兩倍的路徑，但每個路徑的深度是一半。四路徑矩陣乘法加速器 2000 包含用於 Src0 的輸入邏輯 2002A-2002D，用於 Src2 以儲存由輸入邏輯 2010A-2010D 讀取的輸入元素的輸入緩衝區 2011A-2011D，以及用於 Src1 以儲存由共享輸入邏輯 2012 讀取的輸入元素的輸入緩衝區 2013A-2013D。處理元件 2031A-2031B 賦能級 1 的平行處理。處理元件 2032A-2032B 賦能級 2 的平行處理。每個路徑的級 2 可反饋至級 1 或透過輸出邏輯 2022A-2022D 將結果寫入指定的目的地。處理元件 2031A-2031B，2032A-2032B 可包含類似於處理元件 1931A-1931B、1932A-1932B、1931A-1933B、1934A-1934B 的硬體邏輯，並且可使用類似的硬體邏輯來實現回送(loopback)功能。

【0230】二路徑矩陣乘法加速器 1900 或四路徑矩陣乘法加速器 2000 的優點包含可擴縮性(scalability)、軟體相容性及吞吐量。相對於 8 個深度的脈動陣列，這些加速器的模組化架構可實現更有效的擴縮。矩陣乘法加速器的不同配置可適合不同的產品需求或使用案例，而無需重新設計。此外，所使用的相同軟體模型是獨立於硬體實現的。為了由八級的脈動管線執行的指令而設計的演算法可用於使用四級的矩陣乘法加速器的實現。硬體會以一種對軟體透明(transparent)的方式使用反饋來模擬 8 個級的管線。多路徑可用於對 DPAS 指令吞吐量有高要求的設計中。更多路徑的實現可配合更高頻寬的輸入邏輯和輸出邏輯。在一

個實施例中，相對於可能使用的8個深度的脈動陣列，二路徑矩陣乘法加速器1900和四路徑矩陣乘法加速器2000被配置以更高的效率及/或更精細的粒度略過具有區塊稀疏度的輸入。

在可擴縮的矩陣乘法加速器上的稀疏乘法

【0231】第三實施例在處理具有不規則稀疏度的資料時，促進增加的指令吞吐量。透過輸入多工器邏輯可單獨地選擇Src1及Src2輸入的元素，並且只能使用非零值執行處理。

【0232】圖21描繪使用具反饋輸入之脈動陣列的可擴縮的稀疏矩陣乘法加速器2100。可擴縮的稀疏矩陣乘法加速器2100可包含如四路徑矩陣乘法加速器2000中的處理元件2031A-2031D，或者本文所述的任何其他處理元件。在每個路徑的開始處的處理元件2031A-2021B包含用於Src0的輸入邏輯。可擴縮的稀疏矩陣乘法加速器2100的每個路徑的每個級可透過輸入選擇器2112A-2112D接收獨立或共享的Src1的任何元素。每個路徑的每個級亦可接收Src2的任何元素。透過各自的輸入元件選擇器提供獨立的Src2輸入(例如，透過輸入選擇器2110A和輸入選擇器2111A的Src2A，透過輸入選擇器2110B和輸入選擇器2111B的Src2B)。單獨的Src2輸入使不同的路徑能夠計算不同的指令。每個路徑皆有各自的輸出邏輯2122A-2122B，以供輸出不同的指令。

【0233】圖 22 描繪包含稀疏資料之 Src2 輸入 2200A-2200B。在所描繪範例中，稀疏 Src2 輸入 2200A-2200B (Src2A 輸入 2200A[A0, A1, 0, A3, A4, A5, A6, 0]，Src2B 輸入 2200B [B0, 0, B2, B3, 0, 0, 0, 0]) 可在使用共用的 Src1 輸入之可擴縮的稀疏矩陣乘法加速器 2100 的變體的各路徑上進行處理。可擴縮的稀疏矩陣乘法加速器 2100 的每個路徑可接收各自的 Src0 輸入。

【0234】計算過程的第一步係讀取第一個 Src2 元素，並將這些元素重新排列成每一群組 N 個元素，其中 N 係處理這些元素之路徑的深度。具有不同數量的路徑之其他實現可具有不同的群組大小 (group size)。舉例而言，基於四路徑矩陣乘法加速器 2000 的加速器將使用四個資料元素的群組。如果可能的會，只會選擇非零資料元素。舉例而言，Src2A 2200A 的非零值重新排列為三個群組：[A0, A1]、[A3, A4]、[A5, A6]。Src2B 2200B 的非零值重新排列為兩個群組：[B0, B2]、[B3, 0]，其中填充零用以使第二個群組完整。使用此重新排列以便將每一群組的第一元素饋送至路徑的第一級以及將每一群組的第二元素饋送至路徑的第二級。雖然可擴縮的稀疏矩陣乘法加速器 2100 不需要將元素進行分組，但將元素分組會減少可能需要饋送到級的元素數量。

【0235】在計算處理的第二步中，將群組饋送至路徑。替代執行四次傳遞來計算深度為八的指令 (例如，八次反饋傳遞，每一次使用兩個級)，只需三次反饋傳遞，

因為兩個元素為零且不需要處理。反饋的性質允許略過傳遞，將累加器值直接傳送至輸出，而不消耗計算級。為了維持正確的函數計算，將正確的 Src1 元素輸入至給定的 Src2 元素的級以被計算。因此，當處理第二群組 ([A3, A4]) 時，第一級讀取 Src1-3，第二級讀取 Src1-4。當處理 Src2B 2200B 時，只計算兩個群組 ([B0, B2]、[B3, 0])。這兩個群組可使用兩個反饋傳遞來計算，而不是四個。在第一傳遞中，將 Src1-0 和 Src1-2 輸入至第一級和第二級。在第二傳遞中，將 Src1-3 和任何 Src1 元素分別輸入至第一級和第二級。

【0236】針對第三實施例，路徑的深度限制了可被減少的零的數量。在其路徑中具有兩個級的矩陣乘法加速器 (例如，可擴縮的稀疏矩陣乘法加速器 2100) 中，針對八個元素的輸入，只能減少八個、六個、四個及兩個零。在路徑上有四個級的矩陣乘法加速器中，只能減少八個及四個零。為了實現更高解析的稀疏減少，第四實施例在路徑的每一級上增加輸出，並允許每一級接收 Src0 輸入，如圖 23 所示。

【0237】圖 23 顯示使用在每個級中具反饋輸入與輸出之脈動陣列的可擴縮的稀疏矩陣乘法加速器 2300。可擴縮的稀疏矩陣乘法加速器 2300 包含與可擴縮的稀疏矩陣乘法加速器 2100 類似的硬體邏輯，以及附加的輸入和輸出邏輯，以能夠將 Src0 元素提供給每個路徑的每個級並為每個路徑的每個級提供各自的輸出。除了為第一路徑選擇

Src2A元素的輸入選擇器2110A和2111A以及為第二路徑選擇Src2B輸入的輸入選擇器2110A和2111B之外，更針對Src0輸入為每個路徑添加輸入分離器2303A-2303B。每個輸入分離器(splitter)2303A-2303B可包含解多工器或類似的硬體邏輯，以使由輸入邏輯2302A-2302B讀取的Src0輸入元素能夠傳送至每個級。更包含輸入選擇器2112A-2112D，以使每個路徑的每個級選擇Src1輸入。除了來自每個路徑的第二級的輸出邏輯2122A-2122B(處理元件2331C-2331D)之外，更提供了附加的輸出邏輯2322A-2322B，以能夠自每個路徑的第一級(2331A-2331B)提供輸出。另一方面，處理元件2331A-2331C可類似於本文所述的其他處理元件。

【0238】在操作期間，可擴縮的稀疏矩陣乘法加速器2300可被配置以接收只有一個元素的群組。對於第三實施例(例如，可擴縮的稀疏矩陣乘法加速器2100)，Src2上的非零元素分為兩個群組($[B0, B2]$ 、 $[B3, 0]$)，第二群組包含零的填充。圖23中所示的最佳化使得群組可以 $[B0, B2]$ 、 $[B3]$ 形成。 $B0$ 及 $B2$ 將被分配給路徑的第一級和第二級(例如，包含處理元件2331A及處理元件2331C的第一組或包含處理元件2331B和處理元件2331D的第二組中)。於反饋後， $B3$ 將分配給該路徑的第一級。由於路徑的第一級可提供輸出(例如，透過輸出邏輯2322A或2322B)，因此不需要使用路徑的第二級(處理元件2331C或處理元件2331D)。此外，針對該路徑接收的下一個Src2輸入可從第

二級開始，因此一群組的兩個元素將分別分配給第二級和第一級。用於處理新的 Src2 輸入的 Src0 可分配給路徑的第二級(例如，透過輸出邏輯 2322A 或 2322B)。

【0239】除了圖 21 所繪示的可擴縮的稀疏矩陣乘法加速器 2100 和圖 23 所繪示的可擴縮的稀疏矩陣乘法加速器 2300 的硬體邏輯之外，一些實施例額外地包含輸入和輸出硬體記憶體緩衝區。輸入記憶體緩衝區可用於儲存並已存在多個群組的 Src0 和 Src2 輸入，從而減少對高頻寬輸入邏輯的需求。輸出緩衝區使在同一週期內產生的 Dst 輸出能夠以較低的速率穩定地寫入記憶體，從而減少了對高頻寬輸出邏輯的需求。

【0240】另外，一些實施例包含用於所有元素均為零之輸入的略過機制。略過機制允許藉由輸出邏輯直接寫入 Src0，而無需通過脈動陣列。此略過機制與資料相依性策略配合使用，以防止指令之間的寫後讀(read-after-write，RAW)風險可能會破壞資料的完整性。

【0241】圖 24 描繪根據一個實施例的方法 2400，藉由該方法，功能單元處的硬體邏輯可執行指令以執行具累加之脈動點積。方法 2400 可以透過本文所述的可擴縮的稀疏矩陣乘法加速器的硬體及/或韌體邏輯來執行。硬體及/或韌體邏輯可接收非零的源值以及由 GPGPU 的矩陣運算加速器執行的指令的計算深度(2402)。非零的源值可為根據用於可擴縮的稀疏矩陣乘法加速器之路徑的管線深度而分組的非零值。計算深度可指定多個脈動層，以用於計算指令

的點積。邏輯更接收累加器值並將初始值儲存至累加器(2404)。累加器值可為零值(初始累加器值)或由前一個管線級產生的結果。對於指定的計算層，邏輯可評估寫入允許遮罩(write enable mask)以判定一組允許的平行處理通道(2406)。寫入允許遮罩可用於停用特定通道的計算。可基於待執行的指令提供的謂詞遮罩來配置寫入允許遮罩。

【0242】對於每個啟動的平行處理通道，邏輯可基於源輸入元素的逐元素(element-wise)乘法來產生一組乘積(2408)。舉例而言，對於四個元素點積，將兩個源的四個元素相乘以產生該組乘積。在脈動管線的每一層中，相同的Src2元素值係乘以多個不同的Src1值。然後，邏輯可計算該組乘積的總和並將該總和加到累加器中的值(2410)。

【0243】在最後的計算層(2411)上的處理元件上執行方法2400的情況下，處理元件可將計算出的總和輸出至指定的目的地暫存器(2414)。除此之外，處理元件可將其累加器值輸出至下一層(2412)。下一層可為下一實體層或下一虛擬層。輸出到下一個虛擬層包含在處理管線的第一級向處理元件提供反饋值。

【0244】在一個實施例中，圖24的方法2400可由基於以下所示的虛擬碼配置的硬體邏輯來執行。

具累加之四個元素的脈動點積

```
V = Src2.regnum;
temp = Src0.Regnum;    // Accumulated register input
k = Src2.regnum.subregnum;
    for (i = 0; i < sdepth; i++) {
        U = Src1.(Regnum + i);
        Evaluate (WrEn);
        for (n = 0; n < exec_size; n++) {
            if (WrEn.chan[n])
                temp.chan[n] = temp.chan[n]      +
                                U.chan[n].0 * V.k.0 +
                                U.chan[n].1 * V.k.1 +
                                U.chan[n].2 * V.k.2 +
                                U.chan[n].3 * V.k.3;
            //chan[n].0 is a 0th byte in the nth dword
        }
    }
    k ++
}

Dst.regnum = temp;    // Write to output register
```

【0245】在上面顯示的虛擬碼中，Src0、Src1和Src2係儲存運算元資料的暫存器。脈動深度由sdepth指定。執行大小與exec_size對應，並指定平行處理通道的數量。目的地由Dst暫存器指定。在虛擬碼中，識別的暫存器參照於regnum和subregnum欄位。regnum欄位提供運算元的暫存器編號。subregnum欄位提供運算元的子暫存器編號。subregnum欄位與對應的RegNum欄位一同針對暫存器區域的起始點(origin)提供一個位元組對齊位址(byte aligned address)。對於某些指令，此欄位提供位元組位址的位元[4:0]，而RegNum欄位提供位元[12:5]。

【0246】圖25描繪使用稀疏Src2輸入矩陣執行矩陣乘法運算之方法2500。方法2500可透過如本文所述的可擴縮的稀疏矩陣乘法加速器的硬體及/或韌體邏輯來執行。方法2500指定使用稀疏資料的操作，例如圖22之稀疏的Src2輸入2200A-2200B。方法2500可使用圖21的可擴縮的稀疏矩陣乘法加速器2100及/或圖23的可擴縮的稀疏矩陣乘法加速器2300來實現。

【0247】方法2500用於硬體及/或韌體邏輯以將第一矩陣及第二矩陣的多個資料元素讀入矩陣乘法加速器的記憶體中(2502)。接著，該邏輯檢測該第二矩陣的該多個資料元素內的複數非零值(2504)。可以使用矩陣乘法加速器內的向量比較邏輯來執行檢測。接著，該邏輯可將第二矩陣的多個資料元素內檢測到的非零值分組至包含一個或多個資料元素的群組(2506)。接著，該邏輯可提供該群組的資料元素至該矩陣乘法加速器之處理管線的對應級(2508)。對於具有兩個級管線的路徑，群組會包含兩個Src2元素。該群組的第一個元素將提供給第一級，該群組的第二個元素將提供給第二級。針對可擴縮的稀疏矩陣乘法加速器2100，視需要使用零填充來填滿群組。針對可擴縮的稀疏矩陣乘法加速器2300，不需要零填充。

【0248】接著，該邏輯可提供該第一矩陣的該多個資料元素至該處理管線的對應級(2510)。提供的多個資料元素係該些Src1元素，其與待管線級計算之活動的通道相對應。如果元素運作會使用Src2零填充值，則可提供任一

Src1元素。在處理管線之每個活動的級的處理元件執行乘法和累加操作(2512)。在某些情況下，例如，在可擴縮的稀疏矩陣乘法加速器2300正在處理單一元素群組的情況下，對於一指令而言，並非管線的所有級皆為活動的。如果對於該指令某個級不是活動的，則該級仍可用於執行另外的指令之操作。然後，邏輯可從處理管線的每個活動的級輸出或反饋累加值(2514)。當指令之處理的最後一級完成時，輸出可寫入至目的地暫存器或記憶體位置。或者，可將累加值傳送至下一個管線級。輸出至下一個管線級可能涉及輸出至處理管線的第一實體級之寫入反饋(writing feedback)。

附加的示例性計算裝置

【0249】圖26係根據一實施例之包含圖形處理器2604的計算裝置2600之方塊圖。計算裝置2600的版本可為通訊裝置或包含於其中，該通訊裝置諸如機上盒(例如，基於網際網路的有線電視機上盒等)、基於全球定位系統(GPS)的設備位置等等。計算裝置2600亦可為行動計算裝置或包含於其中，行動計算裝置諸如行動電話、智慧型手機、個人數位助理(PDA)、平板電腦、筆記型電腦、電子閱讀器、智慧型電視、電視平台、穿戴式裝置(例如，眼鏡、手錶、手鐲、智慧卡、首飾、衣物等等)及媒體播放器等等。舉例而言，在一個實施例中，計算裝置2600包含採用積體電路(「IC」)(諸如系統單晶片(「SoC」或「SOC」))

的行動計算裝置，整合計算裝置2600的各種硬體及/或軟體組件於單一個晶片上。計算設備2600可為諸如圖1中的資料處理系統100之類的計算裝置。

【0250】計算裝置2600包含圖形處理器2604。圖形處理器2604表示本文所述的任何圖形處理器。在一個實施例中，圖形處理器2604包含快取2614，其可為單個快取或被劃分為快取記憶體的多個區段，包含但不限於任何數量的L1、L2、L3或L4快取、顯像快取、深度快取、取樣器快取及/或著色器單元快取。在一個實施例中，快取2614可為與應用處理器2606共享的最後一階快取。

【0251】在一個實施例中，圖形處理器2604包含圖形微控制器，其實現用於圖形處理器的控制和排程邏輯。控制和排程邏輯可為由圖形微控制器2615執行的韌體。韌體可在啟動時由圖形驅動器邏輯2622加載。韌體亦可被程式化至電子可抹除可程式化唯讀記憶體中，或者自圖形微控制器2615內的快閃記憶體裝置加載。韌體可啟動包含裝置管理/驅動器邏輯2617、2618和排程器2619的GPU OS 2616。GPU OS 2616亦可包含圖形記憶體管理器2620，其可輔助或取代圖形驅動器邏輯2622內的圖形記憶體管理器2621。

【0252】圖形處理器2604更包含GPGPU引擎2644，其包含一個或多個圖形引擎、圖形處理器核心以及本文所述的其他圖形執行資源。這樣的圖形執行資源可以包含但不限於執行單元、著色器引擎、片段處理器、頂點處理器、

串流多處理器、圖形處理器群或適合於處理圖形資源或影像資源或在異質處理器中執行通用計算操作的任何計算資源集合的形式呈現。如圖 11B-11D 所示，GPGPU 引擎 2644 的處理資源可包含在與基板連接的多個硬體邏輯磚中。GPGPU 引擎 2644 可包含複數 GPU 磚 2645，其包含圖形處理與執行資源、快取、取樣器等。舉例而言，如圖 3B-3C 所示，GPU 磚 2645 亦可包含區域揮發性記憶體或可與一個或多個記憶體磚耦接。

【0253】GPGPU 引擎 2644 亦可包含一個或多個特殊磚 2646，其例如包含非揮發性記憶體磚 2656、網路處理器磚 2657 及 / 或通用計算磚 2658。GPGPU 引擎 2644 亦包含矩陣乘法加速器 2660。通用計算磚 2658 亦可包含用於加速矩陣乘法運算的邏輯。非揮發性記憶體磚 2656 可包含複數非揮發性記憶體單元 (memory cell) 及控制器邏輯。非揮發性記憶體磚 2656 的控制器邏輯可由裝置管理 / 驅動器邏輯 2617、2618 其中之一來管理。網路處理器磚 2657 可包含網路處理資源，其耦接到計算裝置 2600 的輸入 / 輸出 (I/O) 源 2610 內的實體介面。網路處理器磚 2657 可由裝置管理 / 驅動器邏輯 2617、2618 其中之一或多者來管理。

【0254】矩陣乘法加速器 2660 係如本文所述的模組化可擴縮的稀疏矩陣乘法加速器。矩陣乘法加速器 2660 可包含多個處理路徑，每個處理路徑包含多個管線級。每個處理路徑可執行不同的指令。在各種實施例中，矩陣乘法加速器 2660 可具有本文所述的多個矩陣乘法加速器中的任一

個或多個的架構特徵。舉例而言，在一個實施例中，矩陣乘法加速器 2660 係具反饋迴路之四個深度的脈動陣列 1700，該反饋迴路可被配置以四個邏輯級的倍數(例如，四個、八個、十二個、十六個等)進行操作。在一個實施例中，矩陣乘法加速器 2660 包含具有四級管線的二路徑矩陣乘法加速器 1900 或具有兩級管線的四路徑矩陣乘法加速器 2000 的一個或多個實例。在一個實施例中，矩陣乘法加速器 2660 包含被配置為可擴縮的稀疏矩陣乘法加速器 2100 或可擴縮的稀疏矩陣乘法加速器 2300 的處理元件。矩陣乘法加速器 2660 可被配置以只對至少 Src2 輸入的非零值進行操作，並且亦可略過 Src1 輸入中存在零值的操作。如果存在區塊稀疏度，則可以略過整個子矩陣的操作。矩陣乘法加速器 2660 亦可包含基於這些實施例的任何組合的任何邏輯。

【0255】如圖所示，在一個實施例中，除了圖形處理器 2604 之外，計算裝置 2600 更可包含任何數量和類型的硬體組件及/或軟體組件，包含但不限於應用處理器 2606、記憶體 2608、輸入/輸出(I/O)源 2610。如依圖 3A 所描繪，應用處理器 2606 可與硬體圖形管線互動，以共享圖形管線功能。處理後的資料儲存在硬體圖形管線的緩衝區中，以及狀態資訊儲存在記憶體 2608 中。可將產生的資料傳輸至顯示控制器，以透過諸如圖 3A 的顯示裝置 318 之類的顯示裝置輸出。顯示裝置可為各種類型(例如陰極射線管(CRT)、薄膜電晶體(TFT)、液晶顯示器(LCD)、有機發光

二極體(OLED)陣列等)，以及可被配置以透過圖形化使用者介面向使用者顯示資訊。

【0256】應用處理器2606可包含一個或多個處理器(諸如圖1之一個或多個處理器102)，以及可為中央處理器(CPU)，其至少部分地用於執行用於計算裝置2600的作業系統(OS)2602。OS 2602可作為計算裝置2600的硬體及/或實體資源與一個或多個使用者之間的介面。OS 2602可包含用於計算裝置2600中的各種硬體裝置的驅動器邏輯。該驅動器邏輯可包含圖形驅動器邏輯2622，其可包含圖10的使用者模式圖形驅動器1026及/或核心模式圖形驅動器1029。圖形驅動器邏輯可包含圖形記憶體管理器2621，以管理用於圖形處理器2604的虛擬記憶體位址空間。

【0257】預期在一些實施例中，圖形處理器2604可作為應用處理器2606的一部分(例如實體CPU封裝的一部分)存在，在這種情況下，記憶體2608的至少一部分可由應用處理器2606與圖形處理器2604共享，儘管記憶體2608的至少一部分可由圖形處理器2604獨占，或者圖形處理器2604可具有單獨的記憶體之儲存。記憶體2608可包含緩衝區(例如，訊框緩衝區)之預先分配的區域；然而，本領域的技術人員應理解，實施例不限於此，並且可以使用任何可存取較低層的圖形管線的記憶體。記憶體2608可包含各種形式的隨機存取記憶體(RAM)(例如，SDRAM、SRAM等)，其包含利用圖形處理器2604來顯像桌面或3D圖形場景的應用。諸如圖1的記憶體控制器116之類的記憶體控制

器集線器可存取記憶體2608中的資料，並將其轉送至圖形處理器2604以進行圖形管線處理。記憶體2608可被供計算裝置2600內的其他組件使用。舉例而言，在軟體程式或應用程式的實現中，從計算裝置2600的各種I/O源2610接收的任何資料(例如，輸入圖形資料)在被一個或多個處理器(例如，應用處理器2606)操作之前可被暫時地佇列於記憶體2608中。類似地，由軟體程式判定應經由該等計算系統介面其中之一從計算裝置2600傳送至外部實體或儲存至內部儲存器元素中的資料通常在傳送或儲存之前被暫時地佇列於記憶體2608中。

【0258】I/O源可包含諸如觸控螢幕、觸摸面板、觸控墊、虛擬或正規鍵盤、虛擬或正規滑鼠、埠、連接器、網路裝置等之類，並且可透過平台控制器集線器130來附接，如參圖1所示。另外，I/O源2610可包含一個或多個I/O設備，其被實現用於從計算裝置2600(例如，網路配接器)傳輸資料及/或傳輸資料至計算裝置2600；或被實現用於計算裝置2600內的大型非揮發性儲存器(例如，SSD/HDD)。包含文數字和其他鍵的使用者輸入裝置可用於將資訊和命令選擇傳達給圖形處理器2604。其他類型的使用者輸入裝置為游標控制，例如滑鼠、軌跡球、觸控螢幕、觸控板或游標方向鍵，以將方向資訊和命令選擇傳達給GPU，並控制游標在顯示裝置上的移動。計算裝置2600可採用攝影機和麥克風陣列以觀察手勢、錄音和錄影以及接收與傳送視覺和音訊命令。

【0259】I/O源2610可包含一個或多個網路介面。網路介面可以包含相關聯的網路處理邏輯及/或與網路處理器磚2657耦接。一個或多個網路介面可提供對區域網路(LAN)、廣域網路(WAN)、都會區域網路(MAN)、個人區域網路(PAN)、藍牙、雲端網路、蜂巢式或行動網路(例如,第3代(3G)、第4代(4G)、第5代(5G)等),內部網路、網際網路等等的存取。網路介面可包含例如具有一個或多個天線的無線網路介面。網路介面更可包含例如有線網路介面,以透過網路電纜與遠端裝置進行通訊,該網路電纜可為例如乙太網路電纜、同軸電纜、光纖電纜、序列電纜或平行電纜。

【0260】網路介面可例如藉由符合IEEE 802.11標準以提供對LAN的存取,及/或無線網路介面可例如藉由符合藍牙標準以提供對個人區域網路的存取。其他無線網路介面及/或協定(包含標準的先前及後續版本)亦可支援。除了透過無線LAN標準的通訊或取而代之外,網路介面可使用例如分時多工存取(TDMA)協定、全球行動通訊系統(GSM)協定、分碼多工存取(CDMA)協定及/或任何其他類型的無線通訊協定來提供無線通訊。

【0261】應當理解,比前述範例裝備更少或更多系統可能對某些實現來說是更合適的。因此,計算裝置2600的配置可根據實現的不同而有所變化,這取決於許多因素,例如價格限制、性能要求、技術改良或其他情況。範例包含(但不限於)行動裝置、個人數位助理、行動計算設

備、智慧型手機、行動電話、手機、單向呼叫器、雙向呼叫器、傳訊裝置(messaging device)、電腦、個人電腦(PC)、桌上型電腦、膝上型電腦、筆記型電腦、手持式電腦、平板電腦、伺服器、伺服器陣列或伺服器場、Web伺服器、網路伺服器、網際網路伺服器、工作站、小型電腦、主架(main frame)電腦、超級電腦、網路設備、Web設備、分散式計算系統、多處理器系統、基於處理器的系統、消費型電子產品、可程式化消費型電子產品、電視、數位電視、機上盒、無線存取點、基地台、用戶站、行動用戶中心、無線電網路控制器、路由器、集線器、閘道器、橋接器、交換器、機器或其組合。

【0262】本文描述一種加速器裝置，其包含主機介面，與主機介面耦接的構造互連以及與構造互連耦接的一個或多個硬體磚，該一個或多個硬體磚包含稀疏矩陣乘法加速硬體，該稀疏矩陣乘法加速硬體包含具反饋輸入的脈動陣列。

【0263】一實施例提供平行處理器，其包含解碼單元，用於將指令解碼為解碼指令，其中，該解碼指令係執行平行點積運算的指令；以及管線脈動點積單元。該管線脈動點積單元被配置以透過脈動處理管線之多管線級，執行該解碼指令。於執行該解碼指令之期間，在第一管線級計算的點積被配置以透過輸出硬體選擇性地寫入至從輸出記憶體與第二管線級其中之一選擇的位置，以及在第三管線級中計算的點積被配置以透過輸出硬體選擇性地寫入至

從該輸出記憶體和該第一管線級其中之一選擇的位置。於另一個實施例中，該解碼指令與第一源運算元及第二源運算元相關聯，該第一源運算元係對儲存第一矩陣的多個資料元素之記憶體的參照(reference)，以及該第二運算元係對儲存第二矩陣的多個資料元素之記憶體的參照。

【0264】一實施例提供一種加速器裝置，其包含主機介面，與主機介面耦接的構造互連以及與構造互連耦接的一個或多個硬體磚。該一個或多個硬體磚包含稀疏矩陣乘法加速硬體，該稀疏矩陣乘法加速硬體包含具反饋輸入的模組化脈動處理陣列。該模組化脈動處理陣列包含一個或多個具有第一數量的管線路徑之處理陣列模組，以及該第一數量的管線路徑具有第二數量的管線級。第一管線級可配置以接收來自最後管線級的反饋輸出。

【0265】一實施例提供一種透過具多級處理管線的硬體矩陣乘法加速器對一組輸入矩陣執行點積運算之方法。該方法包含：透過第一源運算元，將第一矩陣的多個資料元素讀入該硬體矩陣乘法加速器的記憶體中；透過第二源運算元，將第二矩陣的多個資料元素讀入該硬體矩陣乘法加速器的該記憶體中；檢測該第二矩陣的該多個資料元素內的複數非零值；將第二矩陣的多個資料元素內的該等非零值分組成包含一個或多個資料元素的群組，其中該群組之一數量個資料元素與硬體矩陣乘法加速器之多級處理管線中的一數量個級相對應；提供該群組的資料元素至該處理管線的對應級，其係藉由將該資料元素廣播至該對應級

之該處理元件的多個通道；將所提供之該群組的資料元素與該第一矩陣的多個資料元素相乘，以產生一組乘積；對該組乘積求和，並將該組乘積之和累加至累加器值；以及將該累加器值寫入該處理管線的下一級。於另一個實施例中，將該累加器值寫入該處理管線的該下一級包含將管線反饋值寫入該處理管線的第一級。額外地，檢測該第二矩陣的該多個資料元素內的該等非零值可包含檢測該硬體矩陣乘法加速器的該記憶體內的該等非零值。

【0266】根據以上敘述，本領域技術人員可理解，實施例的廣泛技術可以各種形式實現。因此，儘管已結合實施例的特定範例描述實施例，但實施例的真實範圍不應受到如此限制，因為在研究圖式、說明書及以下申請專利範圍之後，本領域技術人員而言可輕易認知到其他修改。

【符號說明】

【0267】

100:處理系統

102:處理器

104:快取記憶體

106:暫存器堆

107:處理器核心

108:圖形處理器

109:指令集

110:介面匯流排

111:顯示裝置
112:加速器
116:記憶體控制器
118:外部圖形處理器
119:外部加速器
120:記憶體裝置
121:指令
122:資料
124:資料儲存裝置
125:觸控感測器
126:無線收發器
128:韌體介面
130:平台控制器集線器
134:網路控制器
140:傳統 I/O 控制器
142:通用串列匯流排 (USB) 控制器
143:鍵盤與滑鼠
144:攝影機
146:音訊控制器
200:處理器
202A~N:處理器核心
204A~N:快取單元
206:共享快取單元
208:圖形處理器

210:系統代理器核心

211:顯示控制器

212:環式互連

213:I/O連接

214:記憶體控制器

216:匯流排控制器單元

218:嵌入式記憶體模組

219:圖形處理器核心

221A:子核心

221B:子核心

221C:子核心

221D:子核心

221E:子核心

221F:子核心

222A:執行單元(EU)陣列

222B:執行單元(EU)陣列

222C:執行單元(EU)陣列

222D:執行單元(EU)陣列

222E:執行單元(EU)陣列

222F:執行單元(EU)陣列

223A:執行緒調度與執行緒間通訊(TD/IC)邏輯

223B:執行緒調度與執行緒間通訊(TD/IC)邏輯

223C:執行緒調度與執行緒間通訊(TD/IC)邏輯

223D:執行緒調度與執行緒間通訊(TD/IC)邏輯

223E:執行緒調度與執行緒間通訊(TD/IC)邏輯

223F:執行緒調度與執行緒間通訊(TD/IC)邏輯

224A:執行單元(EU)陣列

224B:執行單元(EU)陣列

224C:執行單元(EU)陣列

224D:執行單元(EU)陣列

224E:執行單元(EU)陣列

224F:執行單元(EU)陣列

225A:3D取樣器

225B:3D取樣器

225C:3D取樣器

225D:3D取樣器

225E:3D取樣器

225F:3D取樣器

226A:媒體取樣器

226B:媒體取樣器

226C:媒體取樣器

226D:媒體取樣器

226E:媒體取樣器

226F:媒體取樣器

227A:著色器處理器

227B:著色器處理器

227C:著色器處理器

227D:著色器處理器

227E:著色器處理器
227F:著色器處理器
228A:共享區域記憶體(SLM)
228B:共享區域記憶體(SLM)
228C:共享區域記憶體(SLM)
228D:共享區域記憶體(SLM)
228E:共享區域記憶體(SLM)
228F:共享區域記憶體(SLM)
230:固定功能區塊
231:幾何/固定功能管線
232:3D管線
233:圖形微控制器
234:媒體管線
235:共享功能邏輯
236:共享記憶體及/或快取記憶體
237:幾何/固定功能管線
238:附加的固定功能邏輯
239:圖形處理單元(GPU)
240A~N:多核心群組
241:排程器/調度器
242:暫存器堆
243:圖形核心/核心
244:張量核心/核心
245:光線追蹤核心/核心

246:中央處理單元(CPU)

247:L1快取與共享記憶體單元

248:記憶體控制器

249:記憶體

250:輸入/輸出(I/O)電路

251:I/O記憶體管理單元(IOMMU)

252:I/O裝置

253:L2快取

254:L1快取

255:指令快取

256:共享記憶體

257:命令處理器

258:執行緒調度器

260A-N:計算單元

261:向量暫存器

262:純量暫存器

263:向量邏輯單元

264:純量邏輯單元

265:區域共享記憶體

266:程式計數器

267:恆定快取

268:記憶體控制器

269:內部直接記憶體存取(DMA)控制器

270:通用圖形處理核心單元

271:記憶體

272:裝置記憶體

300:圖形處理器

302:顯示控制器

304:複製引擎

306:視訊編解碼器引擎

310A:圖形引擎磚

310B:圖形引擎磚

310C:圖形引擎磚

310D:圖形引擎磚

310:圖形處理引擎(GPE)

312:3D管線

314:記憶體介面

315:3D/媒體子系統

316:媒體管線

318:顯示裝置

320:圖形處理器

322:圖形處理引擎群

323A:磚互連

323B:磚互連

323C:磚互連

323D:磚互連

323E:磚互連

323F:磚互連

324:構造互連

325A:記憶體互連

325B:記憶體互連

325C:記憶體互連

325D:記憶體互連

326A:記憶體

326B:記憶體

326C:記憶體

326D:記憶體

328:主機介面

330:計算加速器

332:計算引擎群

336:L3快取

340A:計算引擎磚

340B:計算引擎磚

340C:計算引擎磚

340D:計算引擎磚

403:命令串流器

410:圖形處理引擎

414:圖形核心陣列

415A:圖形核心

415B:圖形核心

416:共享功能邏輯

418:統一返回緩衝區

420:共享功能邏輯

421:取樣器

422:數學

423:執行緒間通訊(ITC)

425:快取

500:執行緒執行邏輯

502:著色器處理器

504:執行緒調度器

505:光線追蹤器

506:指令快取

507A:執行緒控制邏輯

507B:執行緒控制邏輯

508A-N:執行單元

508:執行單元

509A-N:執行單元

510:取樣器

511:共享區域記憶體

512:資料快取

514:資料埠

522:執行緒仲裁器

524:通用暫存器堆陣列(GRF)

526:架構暫存器堆陣列(ARF)

530:傳送單元

532:分支單元

534:SIMD浮點單元(FPU)

535:SIMD ALU

537:指令提取單元

600:執行單元

601:執行緒控制單元

602:執行緒狀態單元

603:指令提取/預取單元

604:指令解碼單元

606:暫存器堆

607:傳送單元

608:分支單元

610:計算單元

611:ALU單元

612:脈動陣列

613:數學單元

700:圖形處理器指令格式

710:128位元指令格式

712:命令運算碼

714:指令控制欄位

716:執行大小欄位

718:目的地運算元

720:來源運算元 src0

722:來源運算元 src1

724:來源運算元 src2

726:存取/位址模式欄位

730:64位元壓縮指令格式

740:運算碼解碼

742:移動與邏輯運算碼群組

744:流程控制指令群組

746:雜項指令群組

748:平行數學指令群組

750:向量數學群組

800:圖形處理器

802:環形互連

803:命令串流器

805:頂點提取器

807:頂點著色器

811:外殼著色器

813:鑲嵌器

817:域著色器

819:幾何著色器

820:幾何管線

823:串流輸出單元

830:媒體管線

829:剖切器

830:媒體管線

831:執行緒調度器

834:視訊前端

837:媒體引擎
840:顯示引擎
841:2D引擎
843:顯示控制器
850:執行緒執行邏輯
851:L1快取
852A:執行單元
852B:執行單元
854:取樣器
856:資料埠
858:紋理快取
870:顯像輸出管線
873:光柵器與深度測試組件
875:共享L3快取
877:像素操作組件
878:顯像快取
879:深度快取
900:圖形處理器命令格式
902:客戶端
904:命令運算碼
905:子運算碼
906:資料
908:命令大小
910:圖形處理器命令序列

912:管線刷新命令
913:管線選擇命令
914:管線控制命令
916:返回緩衝區狀態命令
920:管線判定
922:3D管線
924:媒體管線
930:3D管線狀態
932:3D基元
934:執行命令
940:媒體管線狀態
942:媒體物件命令
944:執行命令
1000:資料處理系統
1010:3D圖形應用程式
1012:著色器指令
1014:可執行指令
1016:圖形物件
1020:作業系統
1022:圖形API
1024:前端著色器編譯器
1026:使用者模式圖形驅動器
1027:後端著色器編譯器
1028:作業系統核心模式功能

1029:核心模式圖形驅動器
1030:處理器
1032:圖形處理器
1034:通用處理器核心
1050:系統記憶體
1100:IP核心開發系統
1110:軟體模擬
1112:模擬模型
1115:暫存器轉移層級(RTL)設計
1120:硬體模型
1130:設計設備
1140:非揮發性記憶體
1150:有線連接
1160:無線連接
1165:第三方協力廠商製造設備
1170:積體電路封裝總成
1172:硬體邏輯
1173:互連結構
1174:硬體邏輯
1175:記憶體小晶片
1180:基板
1182:橋接器
1183:封裝互連
1185:構造

1187:橋接器

1190:封裝總成

1191:I/O

1192:快取記憶體/快取

1193:硬體邏輯

1194:封裝總成

1195:可互換小晶片

1196:基底小晶片

1197:橋接器互連

1198:基底小晶片

1200:系統單晶片積體電路

1205:應用處理器

1210:圖形處理器

1215:影像處理器

1220:視訊處理器

1225:USB控制器

1230:UART控制器

1235:SPI/SDIO控制器

1240:I²S/I²C控制器

1245:顯示裝置

1250:高解析多媒體介面(HDMI)控制器

1255:行動業界處理器介面(MIPI)顯示器介面

1260:快閃記憶體子系統

1265:記憶體控制器

1270:安全引擎

1305:頂點處理器

1310:圖形處理器

1315A~N:片段處理器

1320A:記憶體管理單元(MMU)

1320B:記憶體管理單元(MMU)

1325A:快取

1325B:快取

1330A:電路互連

1330B:電路互連

1340:圖形處理器

1345:核心間工作管理器

1355A~N:著色器核心

1358:拼貼單元

1400:資料處理系統

1402:處理器/應用處理器

1410:統一記憶體

1412:系統記憶體

1414A:原始碼

1414B:編譯碼

1415:編譯器

1416:運行時間庫

1418:GPGPU記憶體

1420:GPGPU

1421:提取與解碼單元/指令提取與解碼單元

1422:排程器控制器

1423:張量加速器

1424A~N:計算區塊

1425:暫存器

1426:電源與效能模組

1427:快取

1430:L3資料快取

1432:共享區域記憶體

1500:指令管線

1502:接受域磚

1504:輸入體積緩衝區

1505:矩陣運算

1506:輸出緩衝區

1508:脈動陣列/脈動張量陣列

1530:記憶體

1600:脈動陣列

1602:輸入

1610A:輸入

1610B:輸入

1610C:輸入

1610D:輸入

1610E:輸入

1610F:輸入

1610G:輸入

1610H:輸入

1611A:處理元件

1611B:處理元件

1611C:處理元件

1611D:處理元件

1611E:處理元件

1611F:處理元件

1611G:處理元件

1611H:處理元件

1612A:輸入

1612B:輸入

1612C:輸入

1612D:輸入

1612E:輸入

1612F:輸入

1612G:輸入

1612H:輸入

1622:輸入

1700:脈動陣列

1702:讀取

1704:資料選擇器

1706:反饋輸入

1722:寫入

1724:資料選擇器

1726:反饋輸出

1800:時序圖

1802:讀取週期

1804:脈動週期

1806:寫入週期

1850:時序圖

1852:讀取週期

1854:脈動週期

1856:寫入週期

1900:二路徑矩陣乘法加速器

1902A:輸入邏輯

1902B:輸入邏輯

1910A:輸入邏輯

1910B:輸入邏輯

1911A:輸入緩衝區

1911B:輸入緩衝區

1912:共享輸入邏輯

1913A:輸入緩衝區

1913B:輸入緩衝區

1922A:輸出邏輯

1922B:輸出邏輯

1931A:處理元件

1931B:處理元件

1932A:處理元件

1932B:處理元件

1933A:處理元件

1933B:處理元件

1934A:處理元件

1934B:處理元件

2000:四路徑矩陣乘法加速器

2002A:輸入邏輯

2002B:輸入邏輯

2002C:輸入邏輯

2002D:輸入邏輯

2010A:輸入邏輯

2010B:輸入邏輯

2010C:輸入邏輯

2010D:輸入邏輯

2011A:輸入緩衝區

2011B:輸入緩衝區

2011C:輸入緩衝區

2011D:輸入緩衝區

2012:共享輸入邏輯

2013A:輸入緩衝區

2013B:輸入緩衝區

2013C:輸入緩衝區

2013D:輸入緩衝區

2022A:輸出邏輯

2022B:輸出邏輯

2022C:輸出邏輯

2022D:輸出邏輯

2031A:處理元件

2031B:處理元件

2031C:處理元件

2031D:處理元件

2032A:處理元件

2032B:處理元件

2032C:處理元件

2032D:處理元件

2100:可擴縮的稀疏矩陣乘法加速器

2110A:輸入選擇器

2110B:輸入選擇器

2111A:輸入選擇器

2111B:輸入選擇器

2112A:輸入選擇器

2112B:輸入選擇器

2112C:輸入選擇器

2112D:輸入選擇器

2122A:輸出邏輯

2122B:輸出邏輯

2200:Src2輸入

- 2200A:Src2A輸入
- 2200B:Src2B輸入
- 2300:可擴縮的稀疏矩陣乘法加速器
- 2302A:輸入邏輯
- 2302B:輸入邏輯
- 2303A:輸入分離器
- 2303B:輸入分離器
- 2331A:處理元件
- 2331B:處理元件
- 2331C:處理元件
- 2331D:處理元件
- 2322A:輸出邏輯
- 2322B:輸出邏輯
- 2400:方法
- 2402:步驟
- 2404:步驟
- 2406:步驟
- 2408:步驟
- 2410:步驟
- 2411:步驟
- 2412:步驟
- 2414:步驟
- 2500:方法
- 2502:步驟

2504:步 驟

2506:步 驟

2508:步 驟

2510:步 驟

2512:步 驟

2514:步 驟

2600:計 算 裝 置

2602:作 業 系 統

2604:圖 形 處 理 器

2606:應 用 處 理 器

2608:記 憶 體

2610:I/O源

2621:圖 形 記 憶 體 管 理 器

2622:作 業 系 統

2614:快 取

2615:圖 形 微 控 制 器

2616:GPU OS

2617:裝 置 管 理 / 驅 動 器 邏 輯

2618:裝 置 管 理 / 驅 動 器 邏 輯

2619:排 程 器

2620:圖 形 記 憶 體 管 理 器

2644:GPGPU引 擎

2645:GPU磚

2646:特 殊 磚

- 2656:非揮發性記憶體磚
- 2657:網路處理器磚
- 2658:通用計算磚
- 2660:矩陣乘法加速器

【發明申請專利範圍】

【請求項1】一種平行處理器，包含：

解碼單元，用以將指令解碼為解碼指令，其中，該解碼指令係執行平行點積運算的指令；以及

管線脈動點積單元，用以透過脈動處理管線之多管線級，執行該解碼指令，其中，於執行該解碼指令之期間，在第一管線級計算的點積被配置以透過輸出硬體選擇性地寫入至從輸出記憶體與第二管線級其中之一選擇的位置，以及在第三管線級中計算的點積被配置以透過輸出硬體選擇性地寫入至從該輸出記憶體和該第一管線級其中之一選擇的位置。

【請求項2】如請求項1之平行處理器，其中，該解碼指令與第一源運算元及第二源運算元相關聯，該第一源運算元係對儲存第一矩陣的多個資料元素之記憶體的參照(reference)，以及該第二源運算元係對儲存第二矩陣的多個資料元素之記憶體的參照。

【請求項3】如請求項2之平行處理器，更包含硬體電路，用以：

基於該第一源運算元，將該第一矩陣的該多個資料元素讀入至該脈動點積單元的記憶體中；以及

基於該第二源運算元，將該第二矩陣的該多個資料元素讀入至該脈動點積單元的記憶體中。

【請求項4】如請求項3之平行處理器，更包含硬體電路，用以檢測該第二矩陣的該多個資料元素內的非零值，

並將該第二矩陣的該多個資料元素內檢測到的非零值分組至包含一個或多個資料元素的群組中。

【請求項5】如請求項4之平行處理器，其中，該群組包含與該脈動處理管線之一數量個管線級相對應的一數量個資料元素。

【請求項6】如請求項5之平行處理器，更包含硬體電路，用以：

提供該群組中的第一資料元素至該第一管線級；以及
提供該群組中的第二資料元素至該第二管線級。

【請求項7】如請求項6之平行處理器，其中，該第三管線級係該第二管線級，以及該脈動處理管線包含二個管線級。

【請求項8】如請求項6之平行處理器，更包含在該第二管線級與該第三管線級之間的第四管線級，其中該脈動處理管線包含四個管線級。

【請求項9】如請求項6之平行處理器，更包含硬體電路，用以：

提供該第一矩陣的第一組多個資料元素至該第一管線級；以及

提供該第二矩陣的第二組多個資料元素至該第二管線級。

【請求項10】如請求項1至9中任一項之平行處理器，其中，該脈動處理管線係第一脈動處理管線，以及該管線脈動點積單元包含第二脈動處理管線。

【請求項 11】如請求項 10 之平行處理器，其中，該第一脈動處理管線和該第二脈動處理管線的各管線級包含一個或多個處理元件，各該處理元件包含多個平行處理器通道，並且各該平行處理器通道可配置以對多個資料元素的向量執行操作。

【請求項 12】一種加速器裝置，包含：

主機介面；

與主機介面耦接的構造互連(fabric interconnect)；以及

一個或多個與該構造互連耦接的硬體磚(hardware tile)，該一個或多個硬體磚包含稀疏矩陣乘法加速硬體，該稀疏矩陣乘法加速硬體包含具反饋輸入之模組化脈動處理陣列，該模組化脈動處理陣列包括多個處理陣列模組，該多個處理陣列模組包含一個或多個具有第一數量的管線路徑之處理陣列模組，該第一數量的管線路徑具有第二數量的管線級，其中，第一管線級可配置以接收來自最終管線級的反饋輸出，且該一個或多個處理陣列模組包含複數管線路徑，該等管線路徑被配置具有用以讀取與第一源輸入相關聯的複數資料元素之共享的硬體電路，以及用以讀取與第二源輸入相關聯的複數資料元素之獨立的硬體電路。

【請求項 13】如請求項 12 之加速器裝置，其中，該一個或多個處理陣列模組包含複數管線路徑，該等管線路徑被配置具有用以讀取與第一源輸入相關聯的複數資料元素

之獨立的硬體電路，以及用以讀取與第二源輸入相關聯的複數資料元素之獨立的硬體電路。

【請求項 14】如請求項 13 之加速器裝置，其中，該一個或多個處理陣列模組包含硬體電路，其被配置以檢測該第二源輸入中的複數非零資料元素，並基於該第二源輸入中的該等非零資料元素及與該第二源輸入的該等非零資料元素相對應之該第一源輸入的複數資料元素，選擇性地執行點積運算。

【請求項 15】如請求項 14 之加速器裝置，其中，該一個或多個處理陣列模組包含管線路徑，該管線路徑包含用於各管線級之獨立的輸出硬體。

【請求項 16】如請求項 15 之加速器裝置，其中，該一個或多個處理陣列模組包含可配置以執行具有第一組輸入的第一點積指令之第一管線路徑以及可配置以執行具有第二組輸入的第二點積指令之第二管線路徑。

【請求項 17】一種執行平行處理之方法，該方法包含：

透過平行處理器的解碼單元解碼指令，該指令被解碼為解碼指令，其中，該解碼指令係執行平行點積運算的指令；以及

透過脈動處理管線之多管線級執行該解碼指令，其中，執行該解碼指令包括：

在第一管線級計算第一點積；

透過輸出硬體將該第一點積選擇性地寫入至從輸

出記憶體與第二管線級其中之一選擇的位置；

在第三管線級中計算第二點積；以及

透過輸出硬體將該第二點積選擇性地寫入至從該輸出記憶體和該第一管線級其中之一選擇的位置。

【請求項 18】如請求項 17 之方法，其中，該解碼指令與第一源運算元及第二源運算元相關聯，該第一源運算元係對儲存第一矩陣的多個資料元素之記憶體的參照 (reference)，以及該第二源運算元係對儲存第二矩陣的多個資料元素之記憶體的參照，且該方法還包含：

基於該第一源運算元，將該第一矩陣的該多個資料元素讀入至該脈動點積單元的記憶體中；以及

基於該第二源運算元，將該第二矩陣的該多個資料元素讀入至該脈動點積單元的記憶體中。

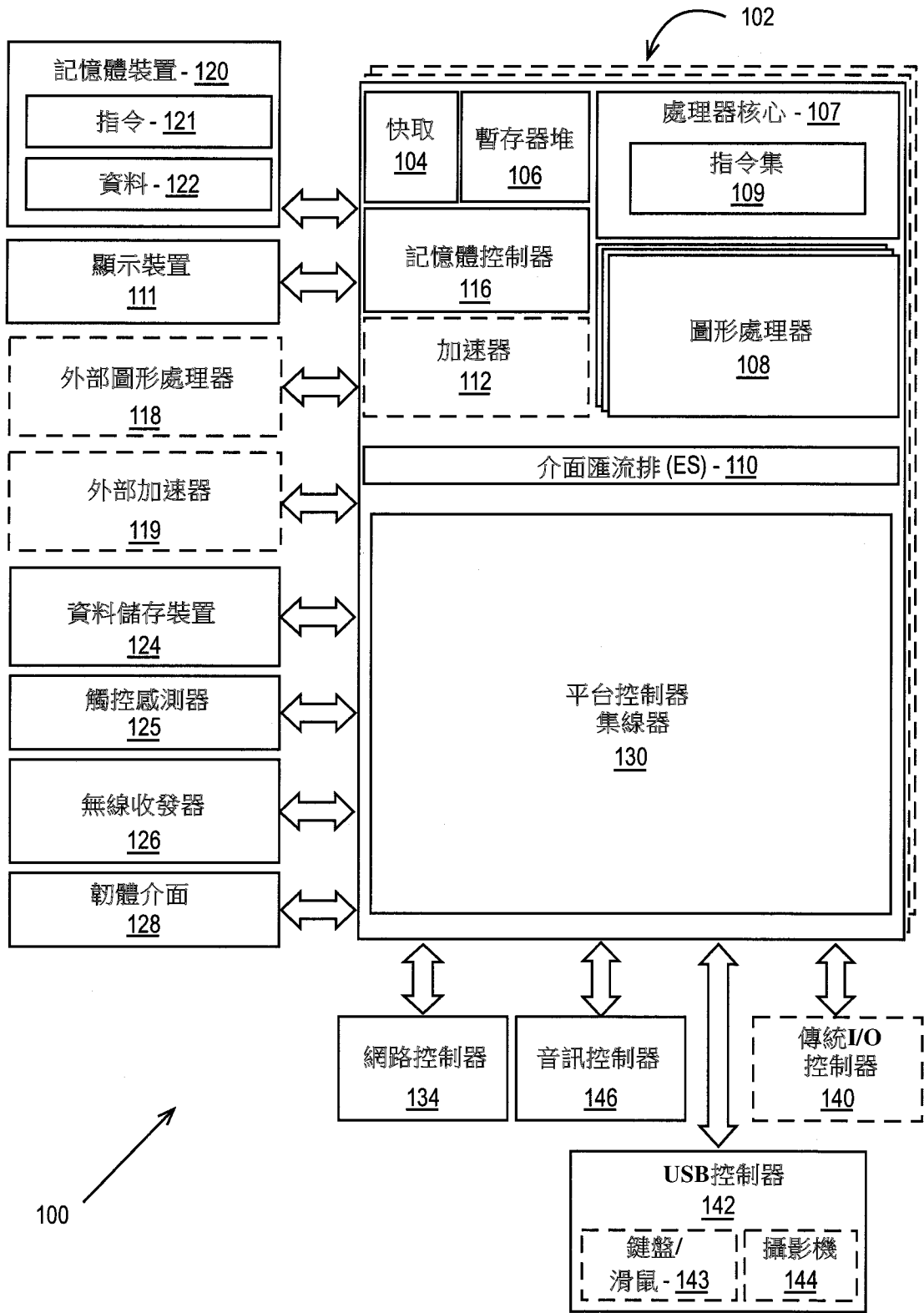
【請求項 19】如請求項 18 之方法，還包含：

透過硬體電路，檢測該第二矩陣的該多個資料元素內的非零值；以及

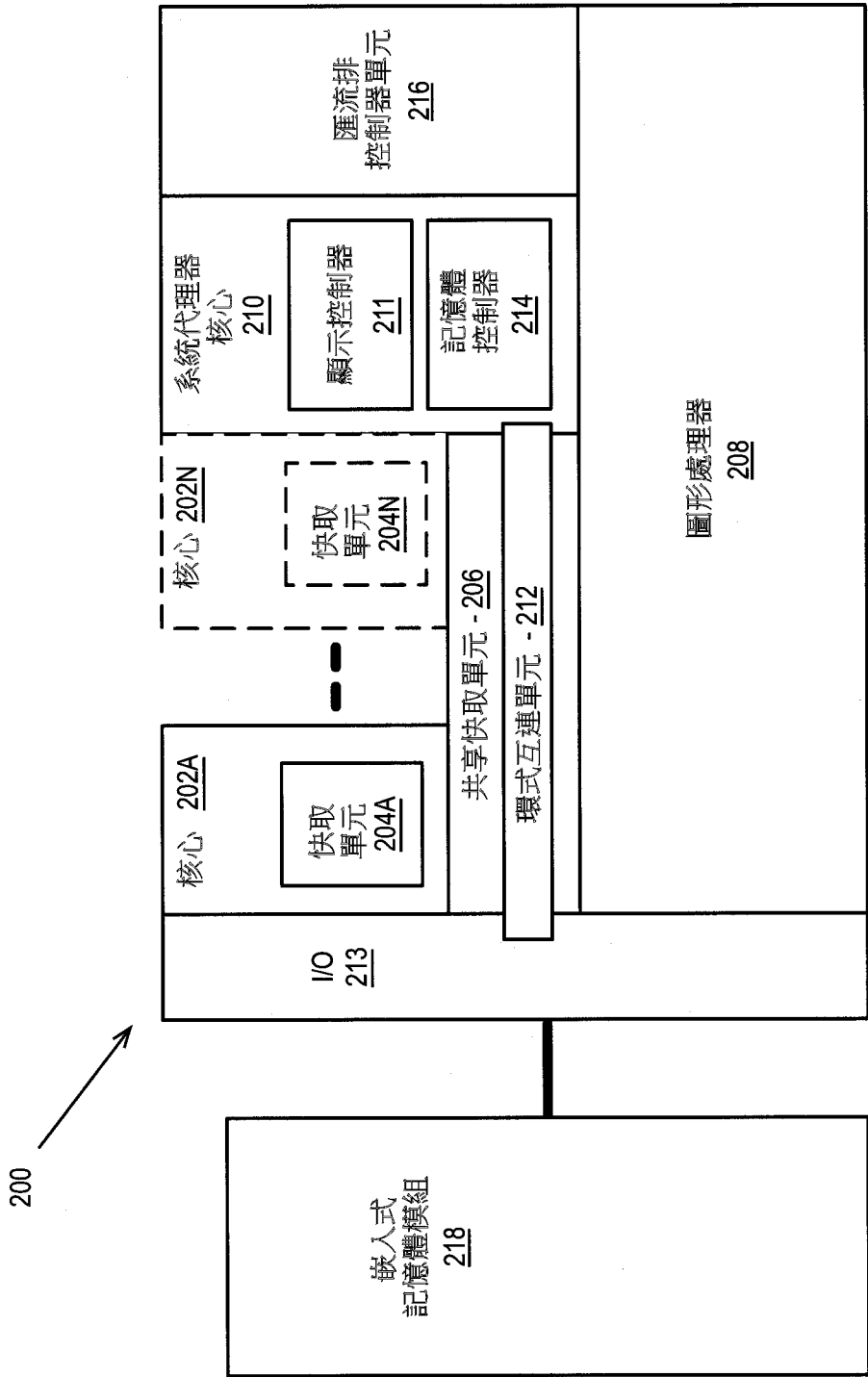
將該第二矩陣的該多個資料元素內檢測到的非零值分組至群組中，該群組包含與該脈動處理管線之一數量個管線級相對應的一數量個資料元素。

【請求項 20】如請求項 19 之方法，還包含提供該群組中的第一資料元素至該第一管線級以及提供該群組中的第二資料元素至該第二管線級。

【發明圖式】

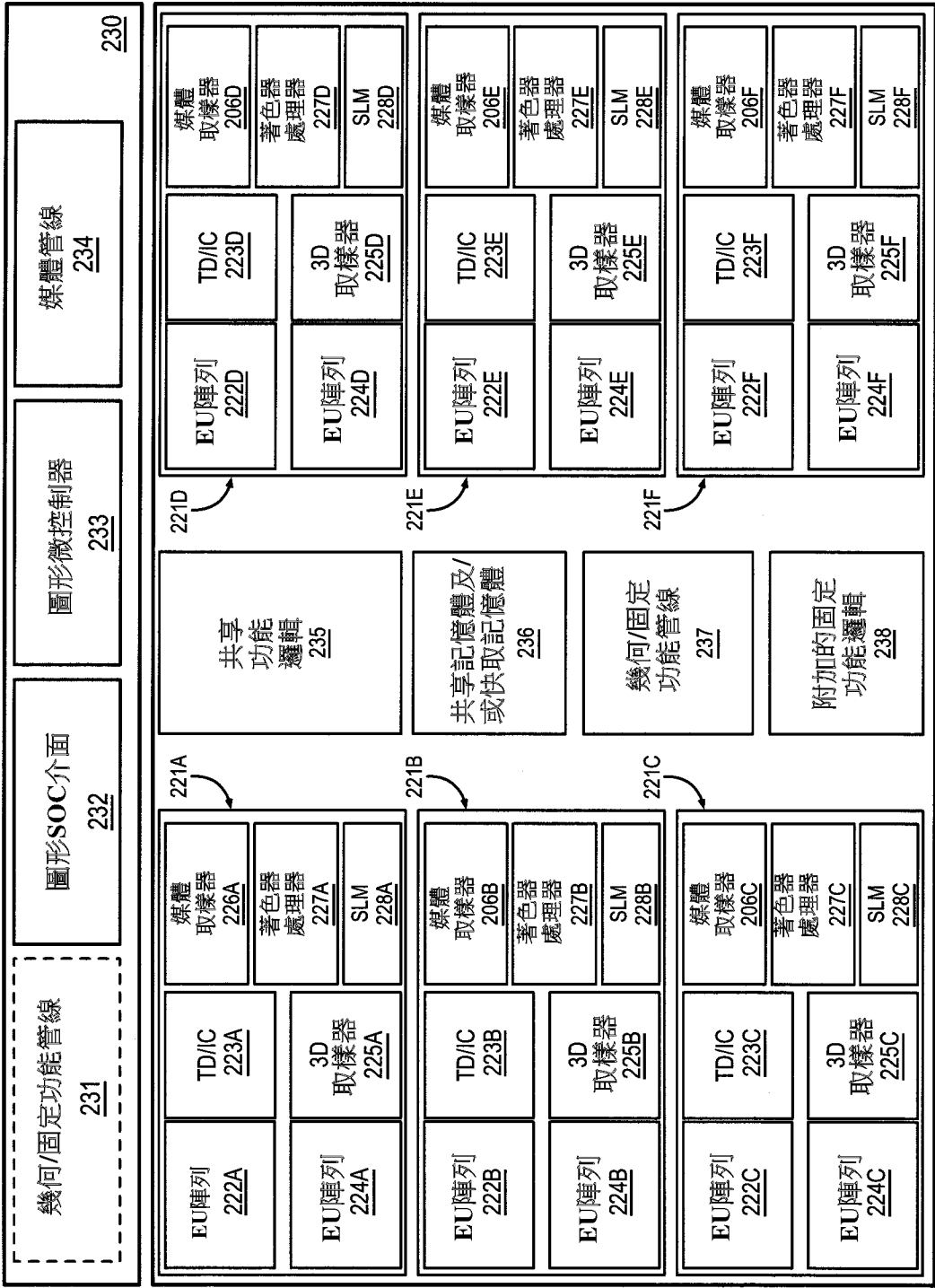


【圖 1】

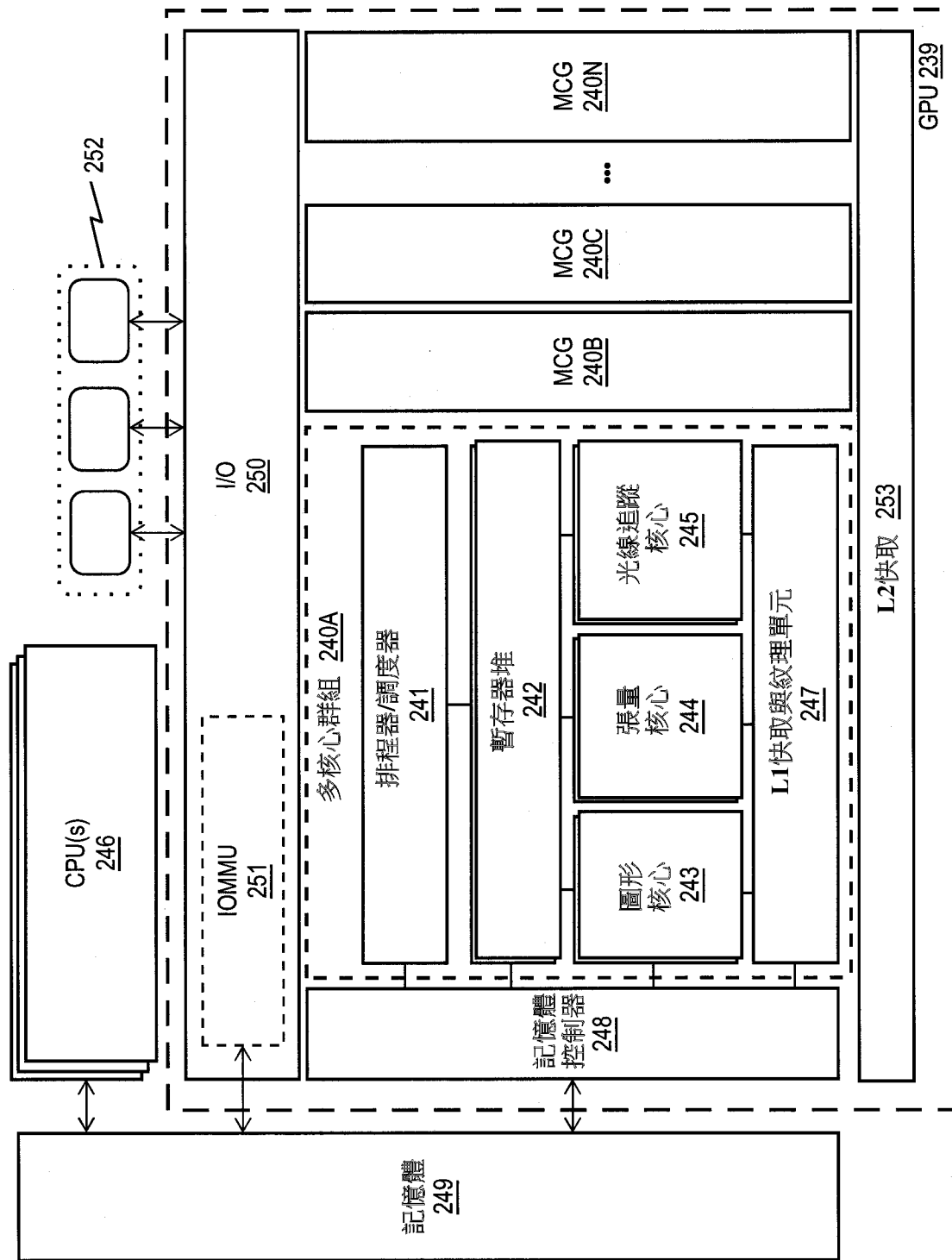


【圖 2A】

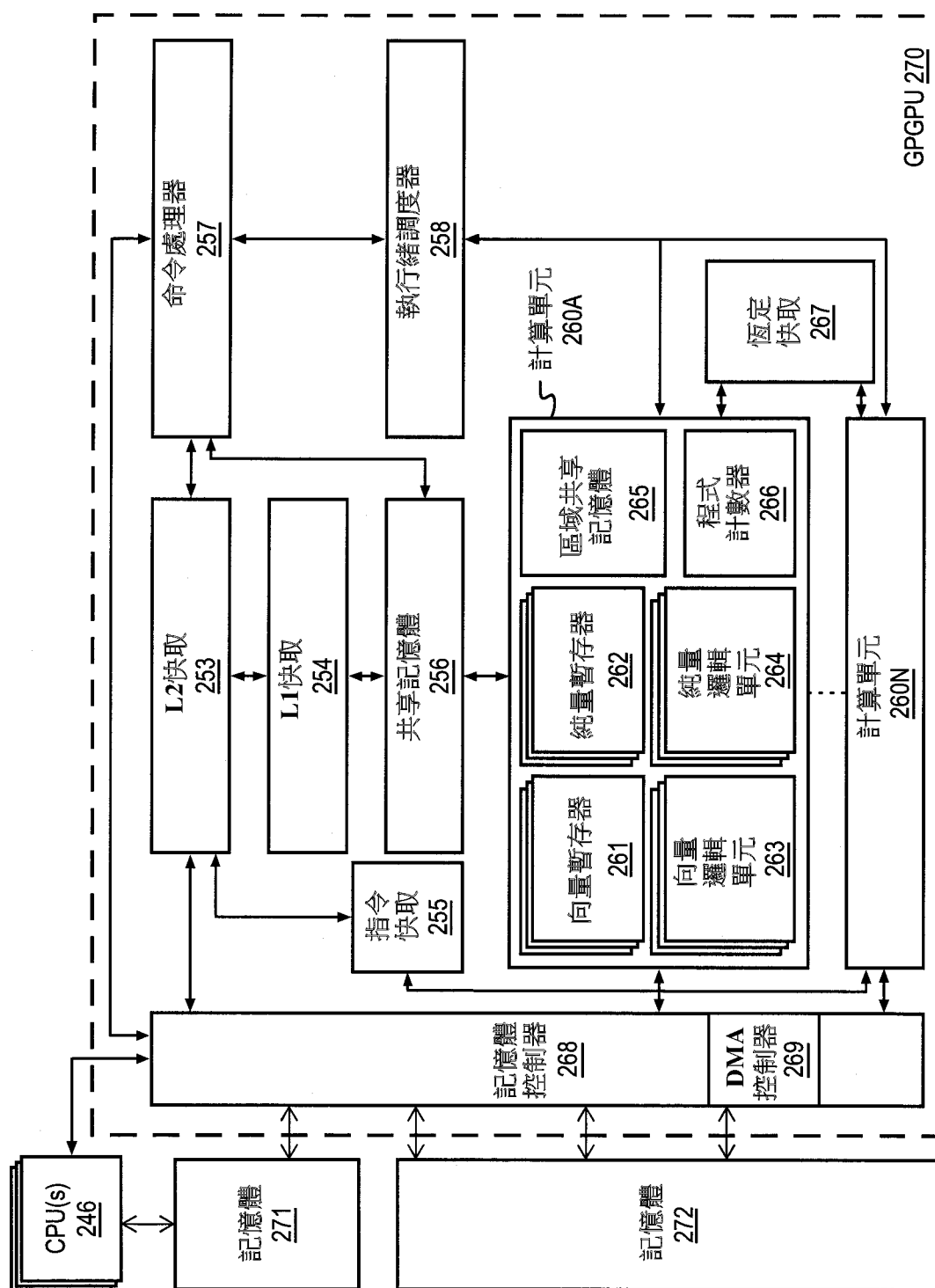
219



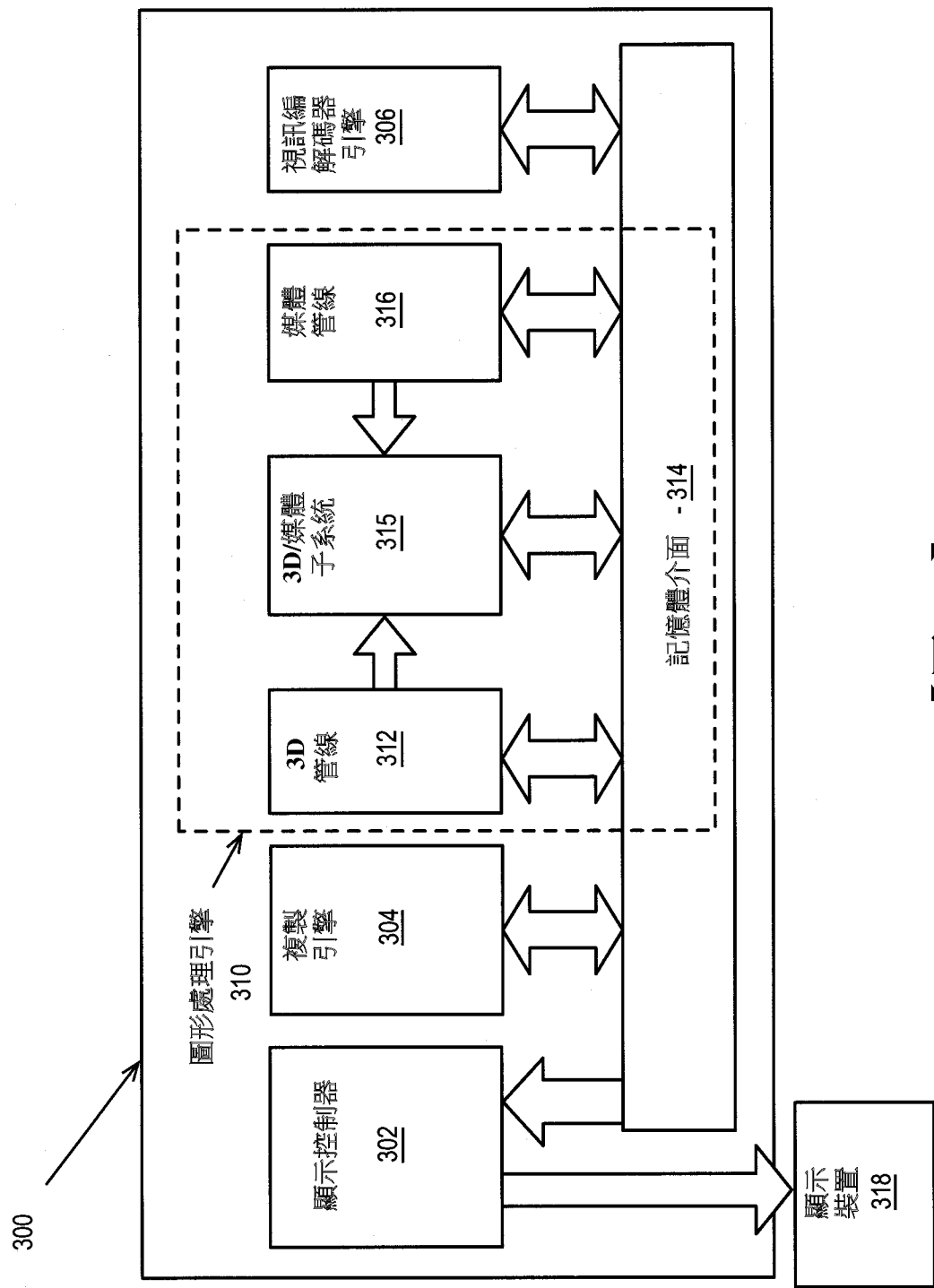
【圖 2B】



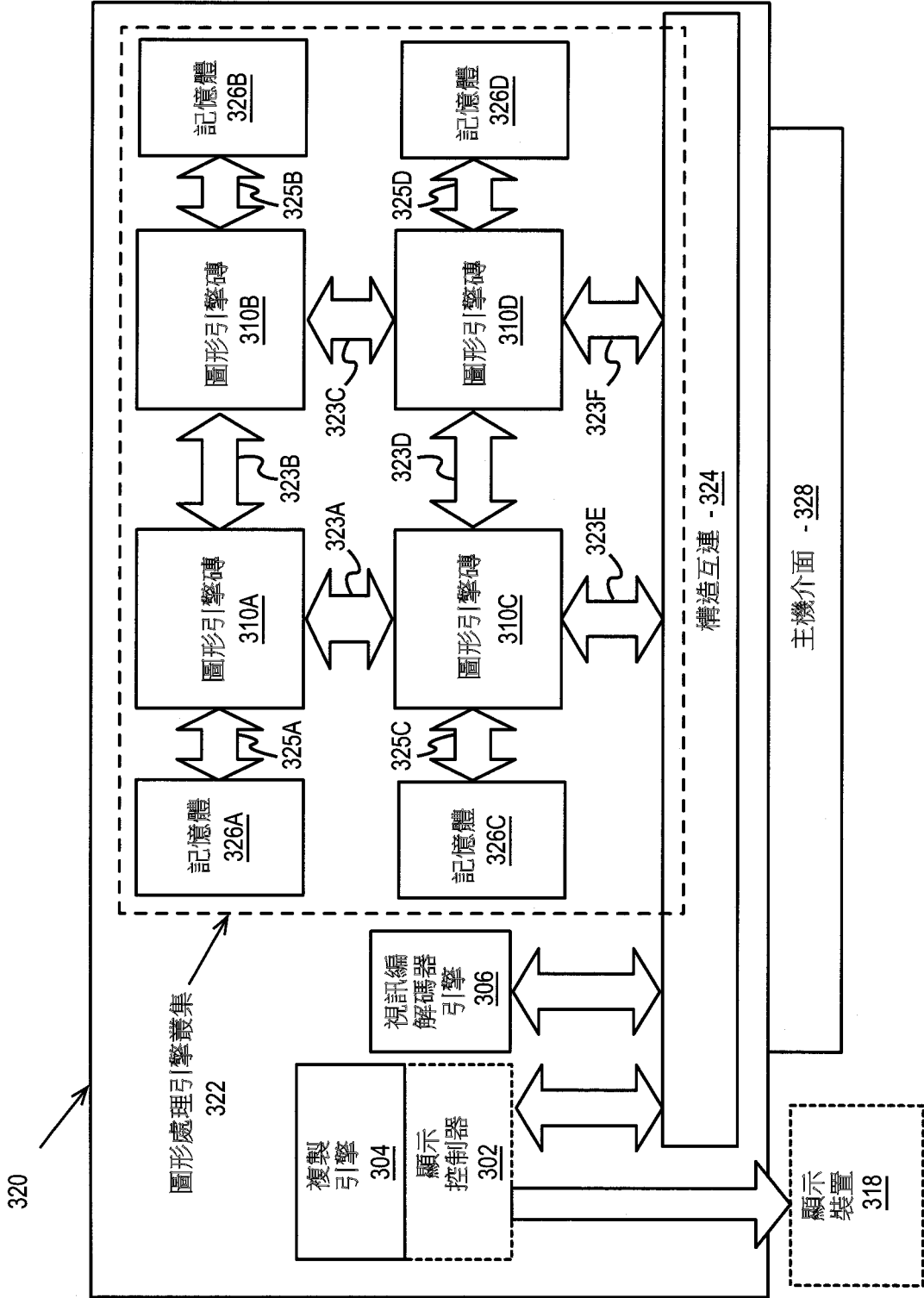
【圖 2C】



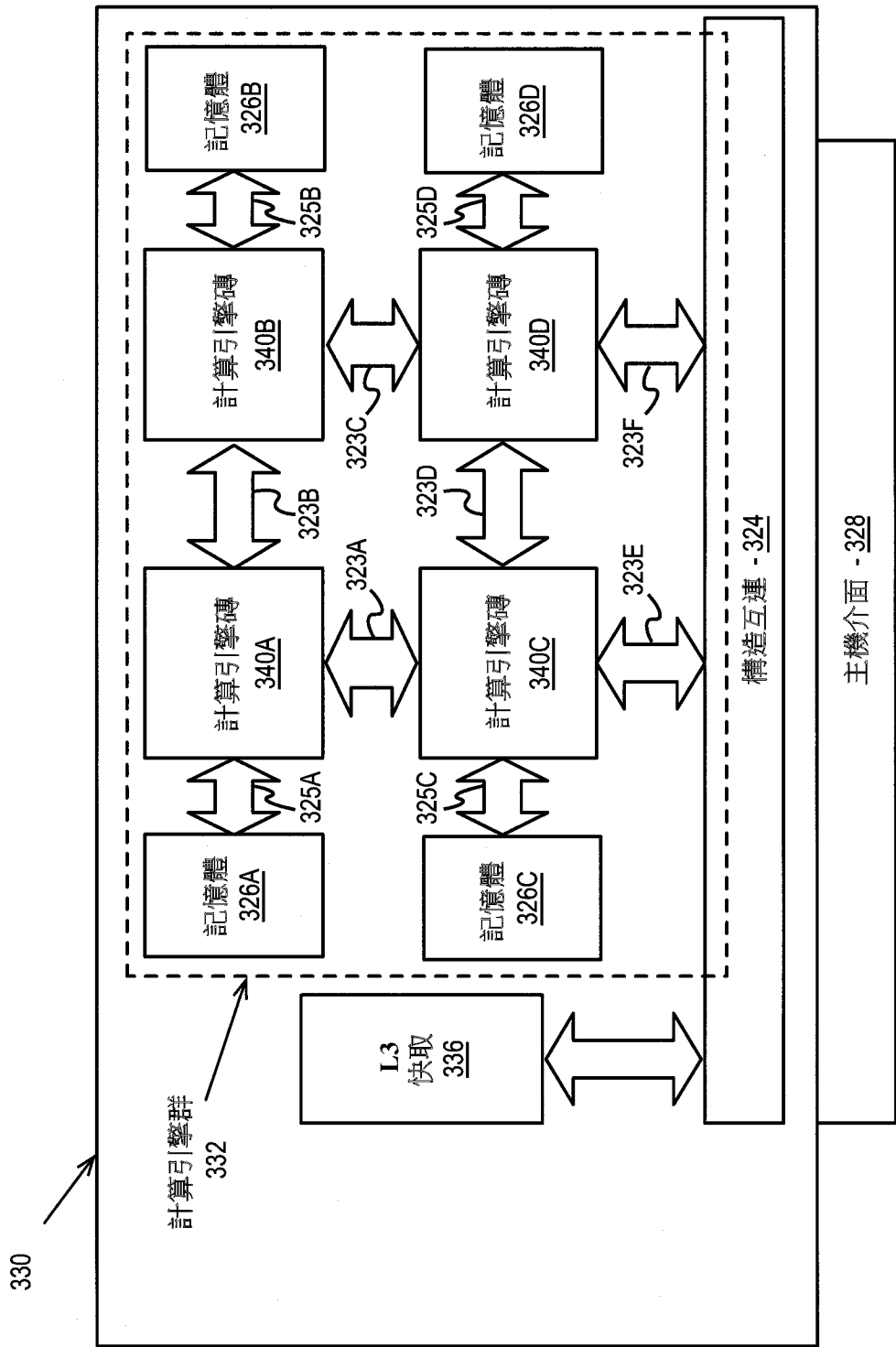
【圖 2D】



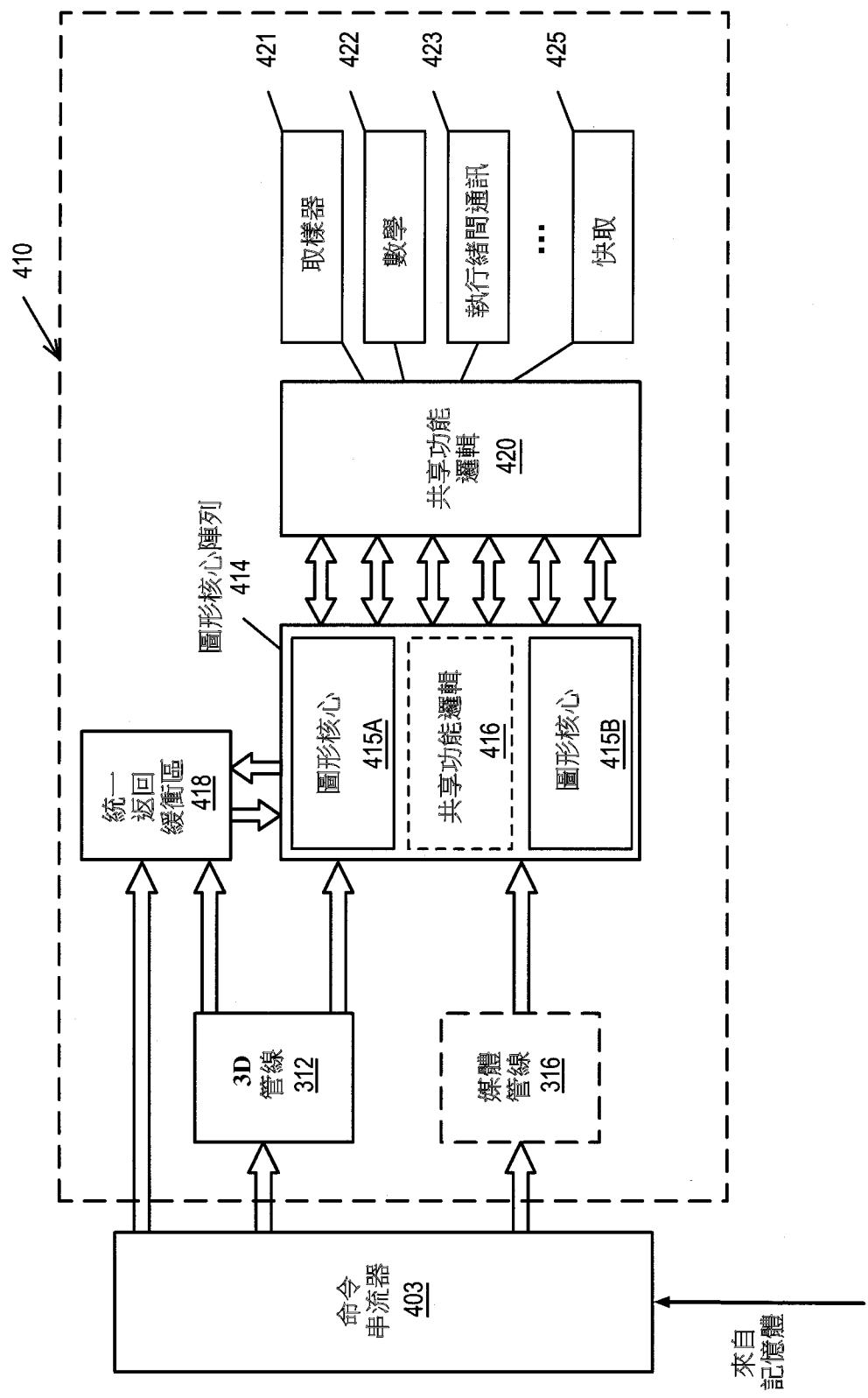
【圖 3A】



【圖 3B】



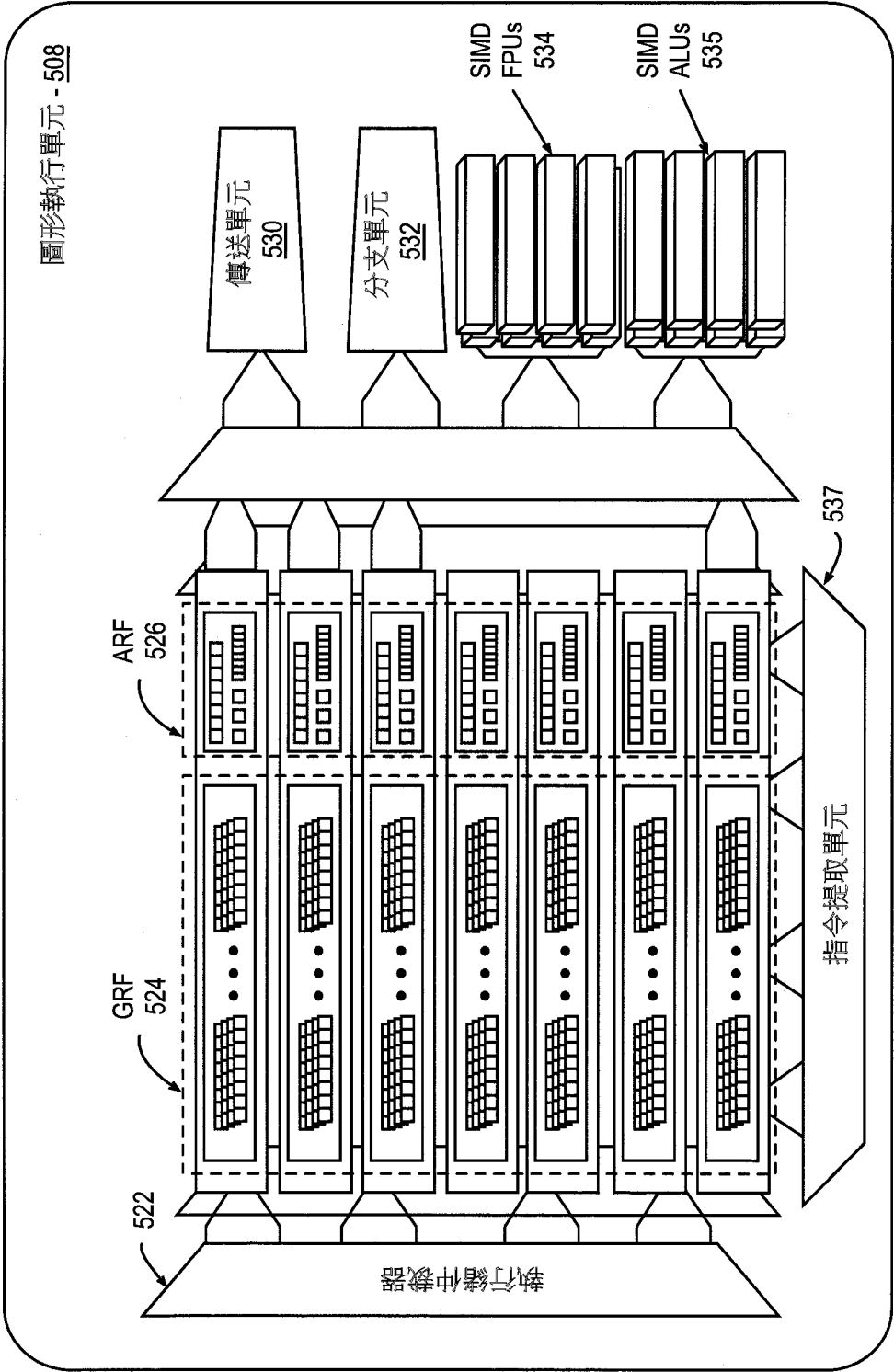
【圖 3C】



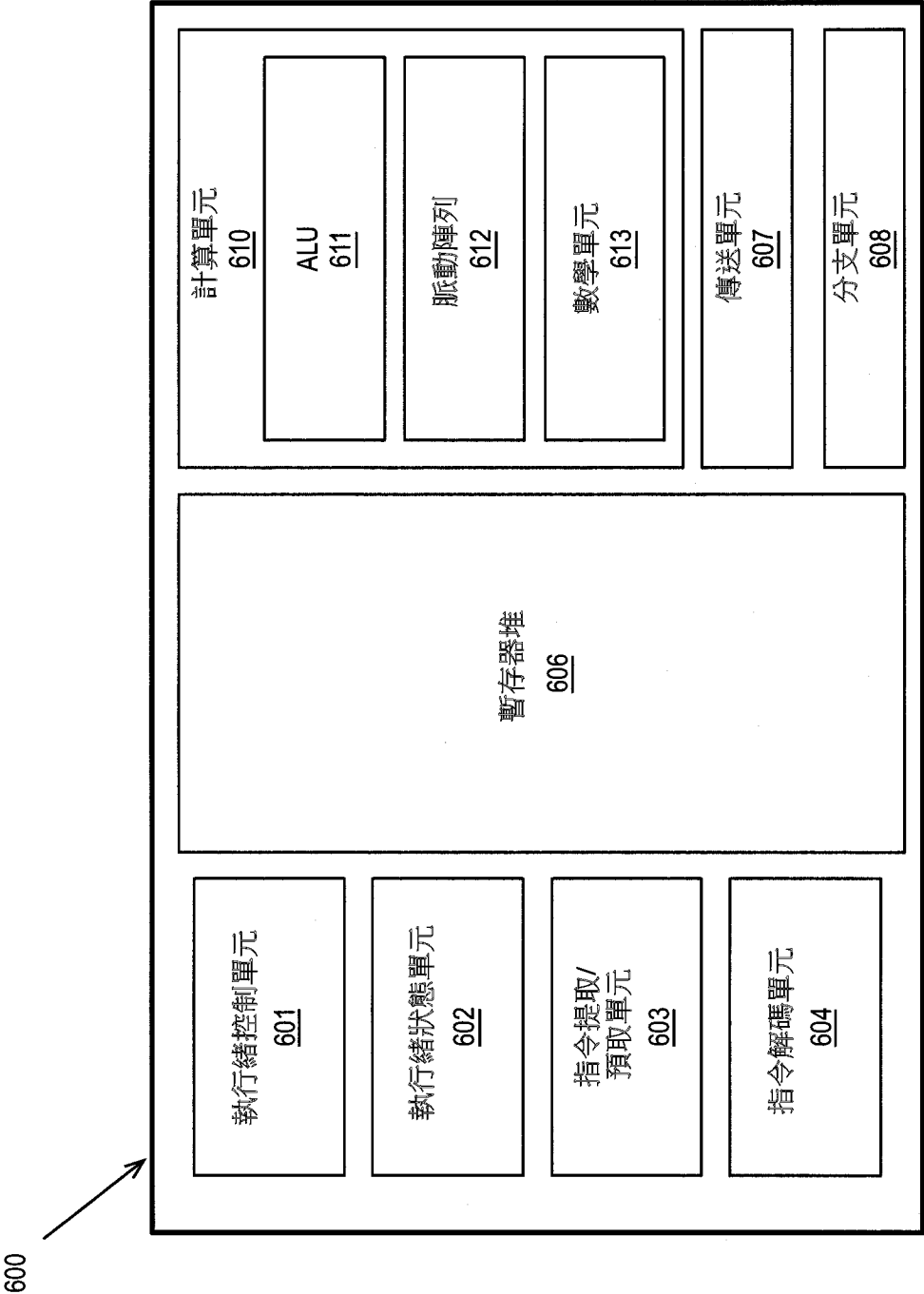
【圖 4】



【圖 5A】



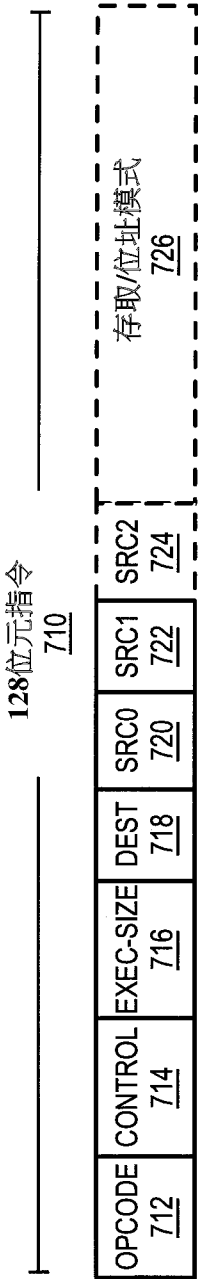
【圖 5B】



【圖 6】

圖形處理器指令格式

700



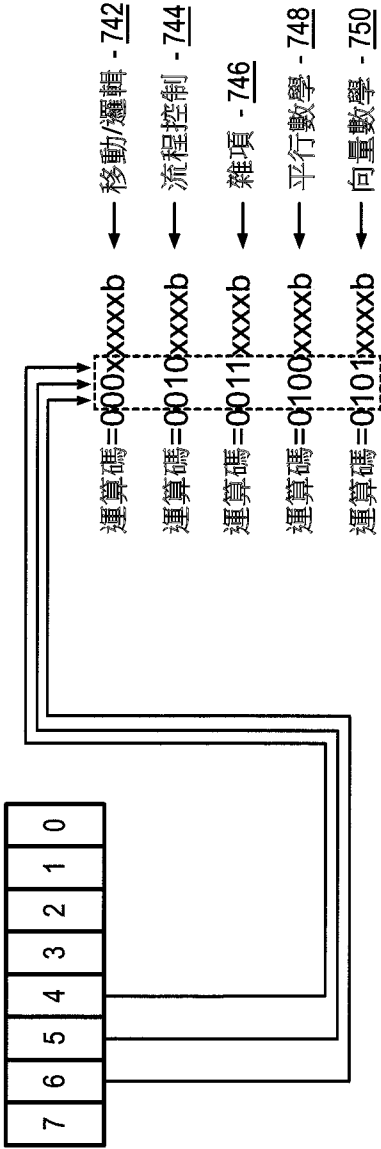
64位元壓縮指令

730

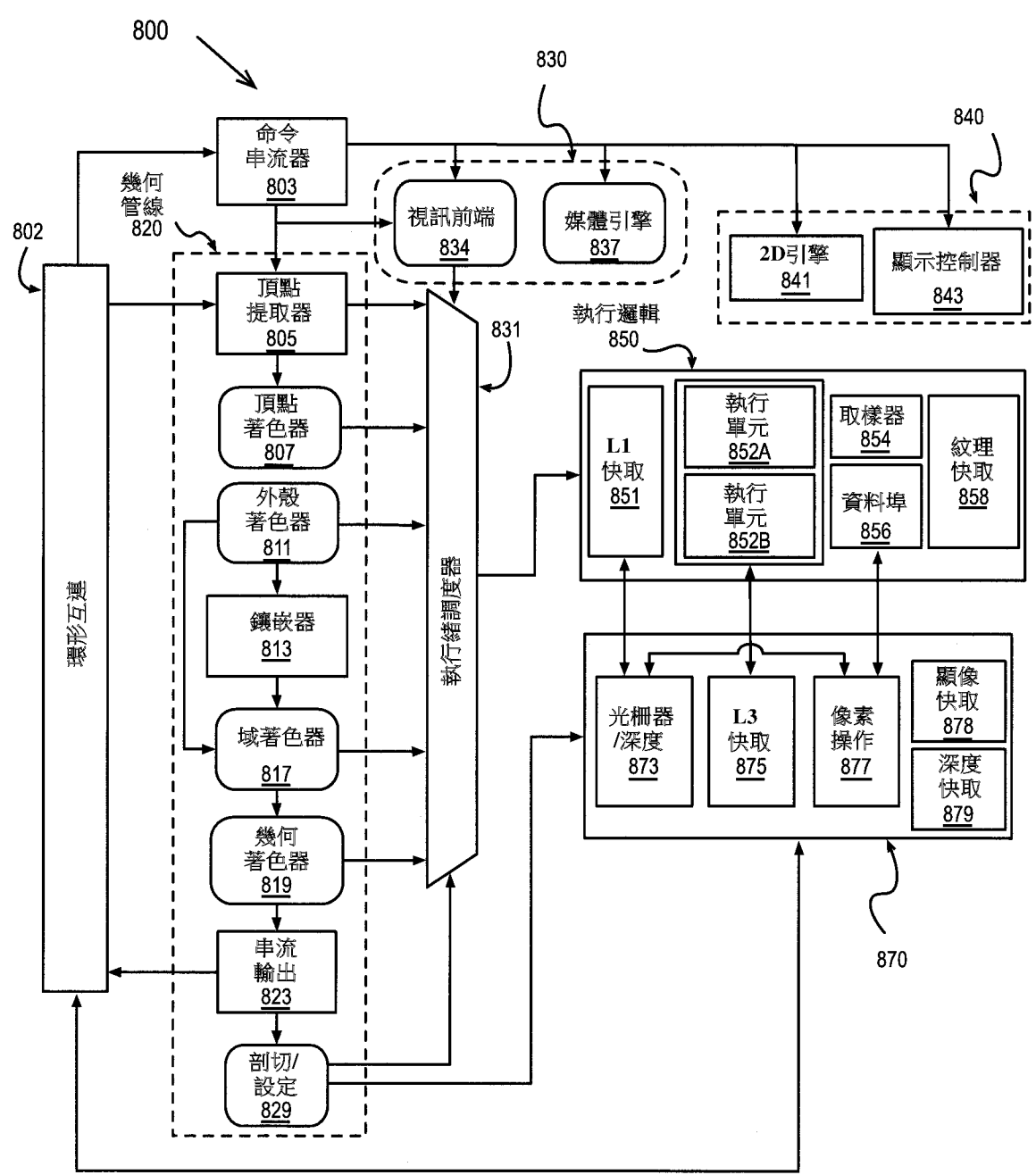


運算碼解碼

740

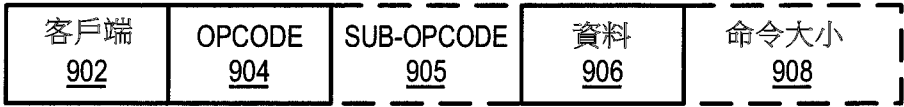


【圖 7】



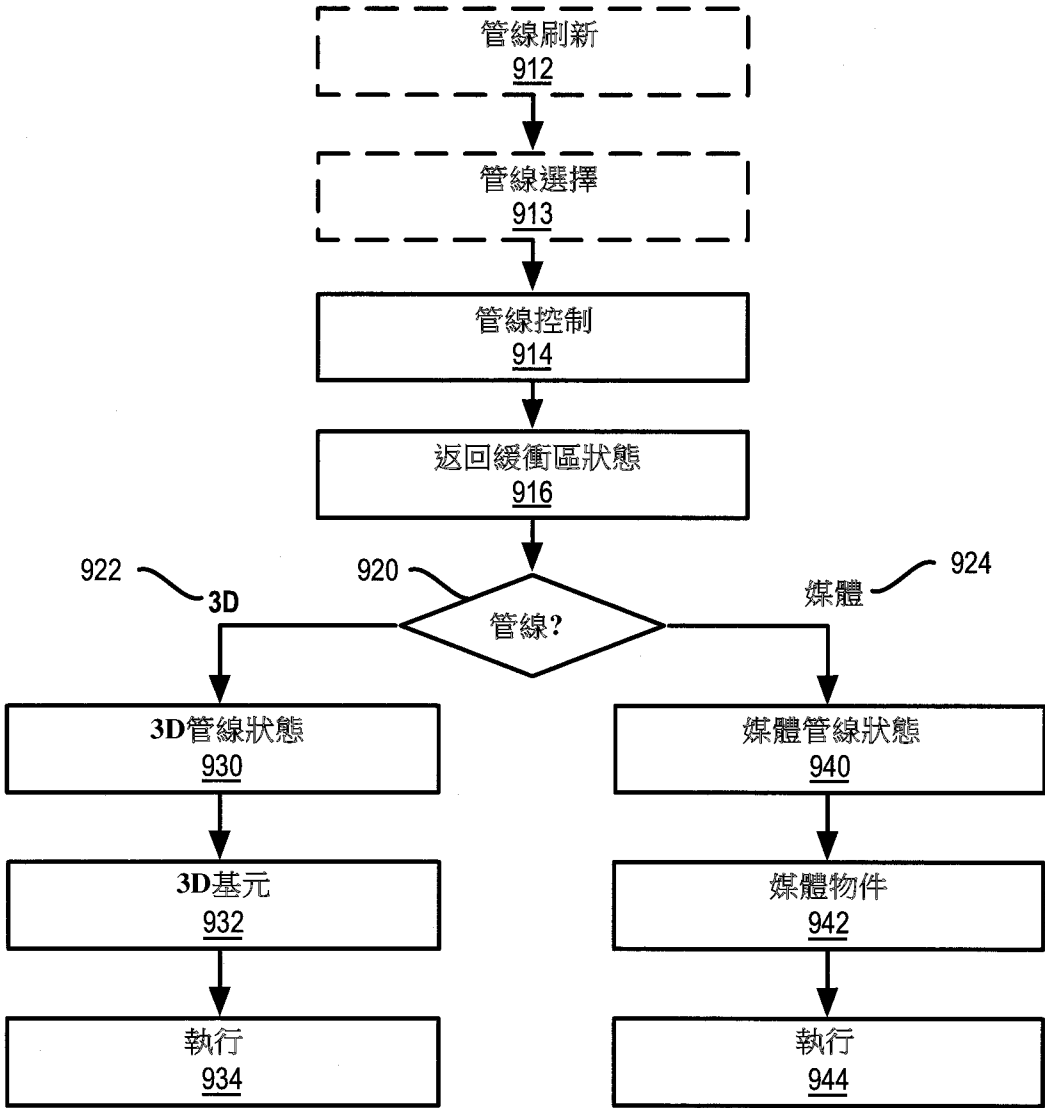
【圖 8】

圖形處理器命令格式
900



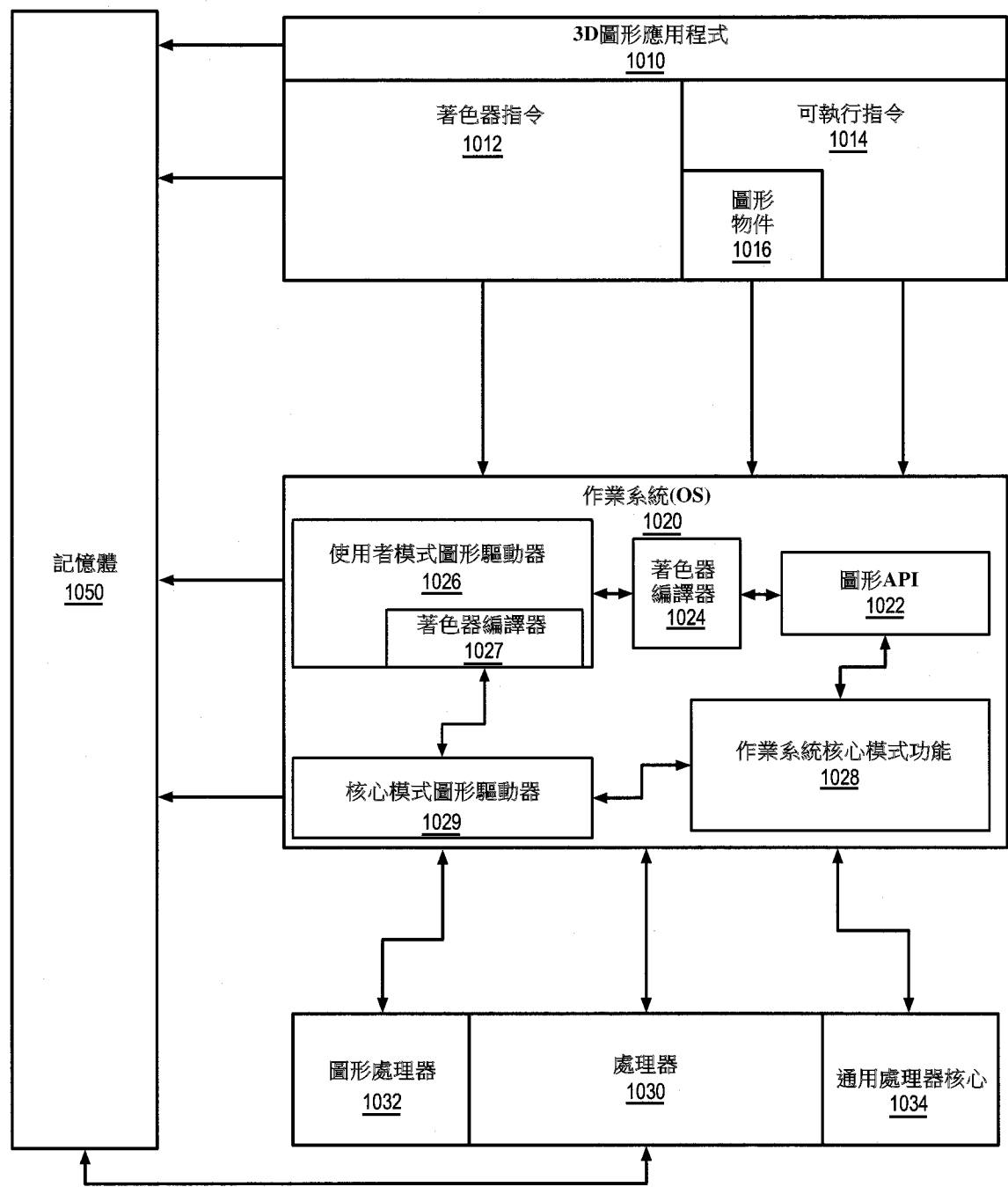
【圖 9A】

圖形處理器命令序列
910



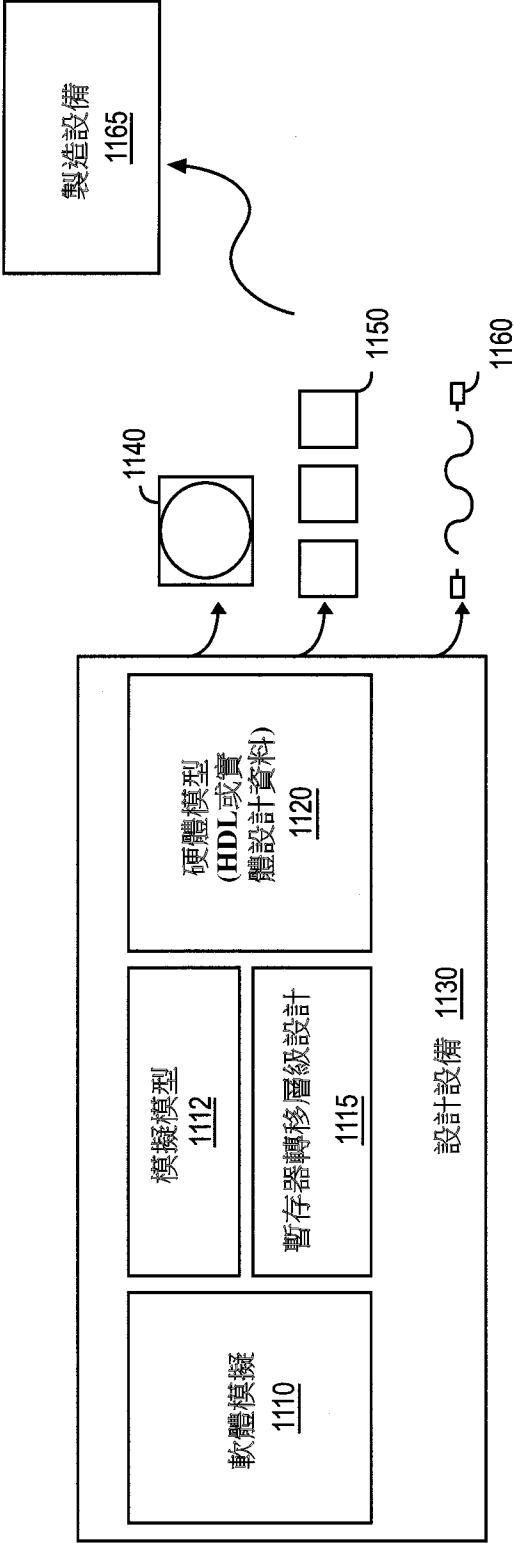
【圖 9B】

資料處理系統 -1000

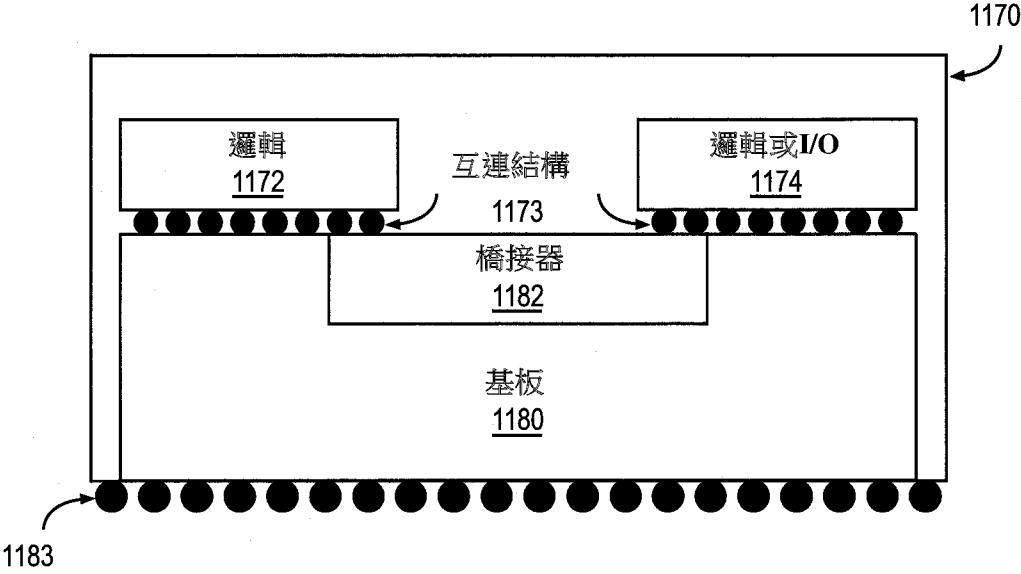


【圖 10】

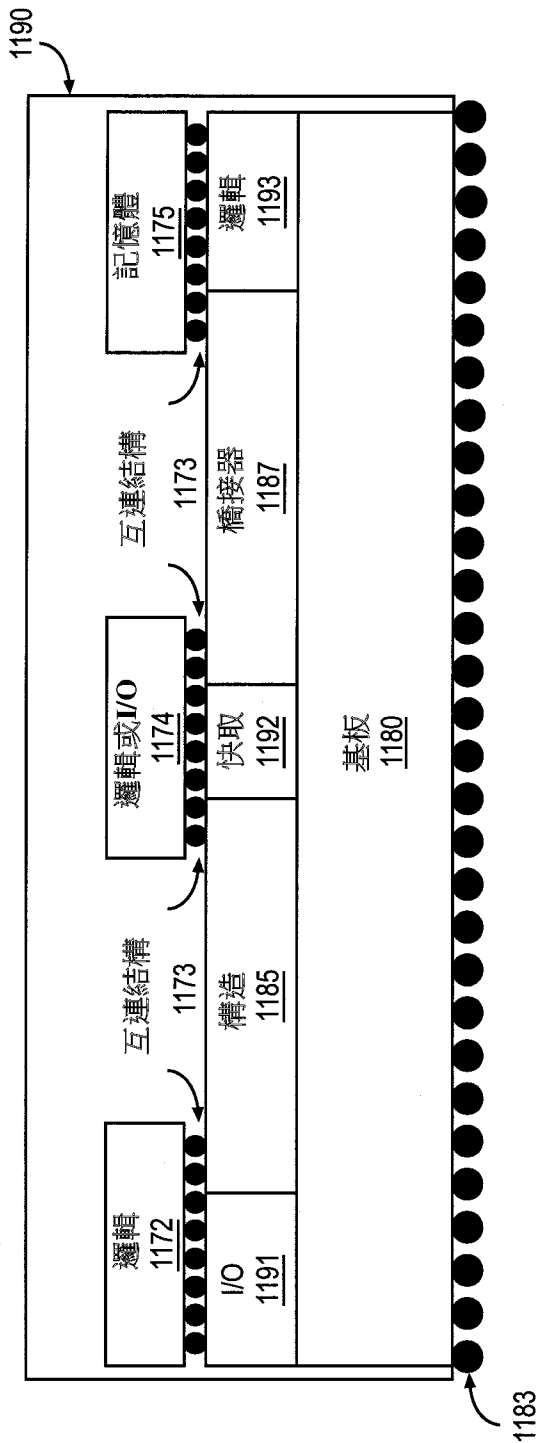
IP核心開發 - 1100



【圖 11A】

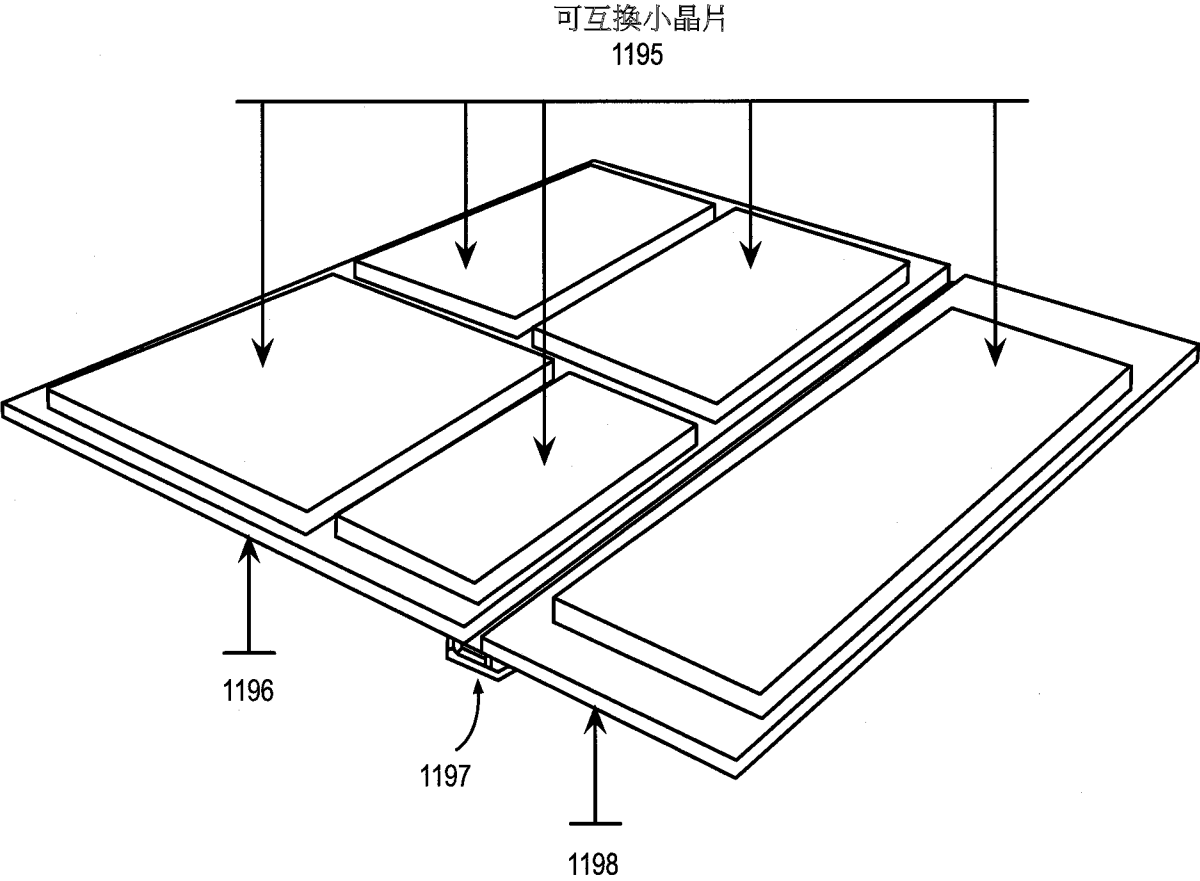


【圖 11B】

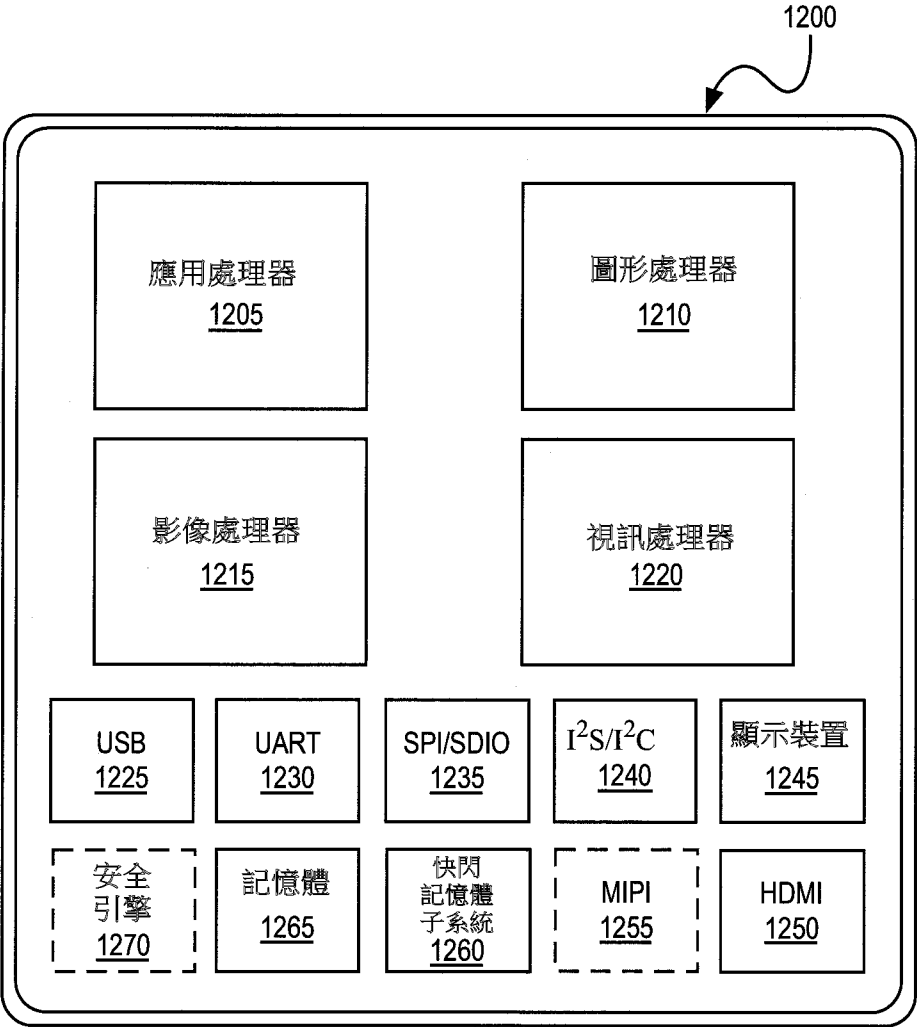


【圖 11C】

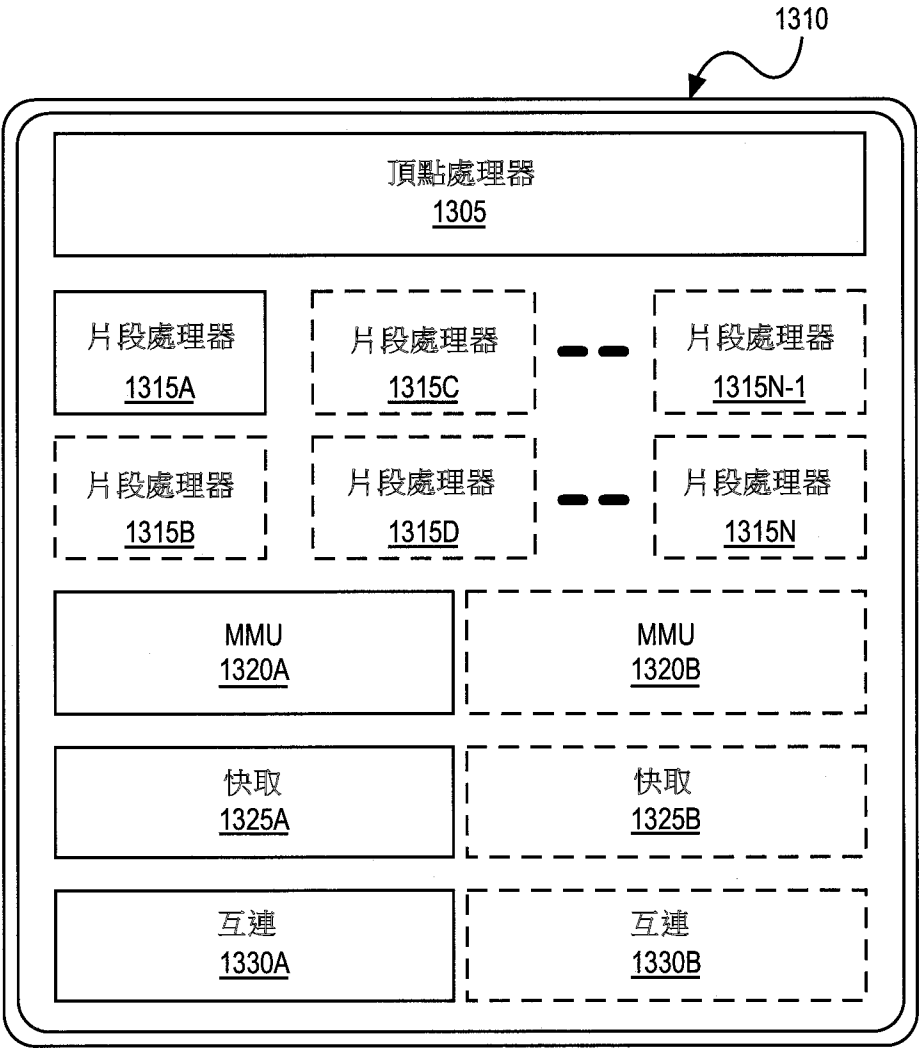
1194



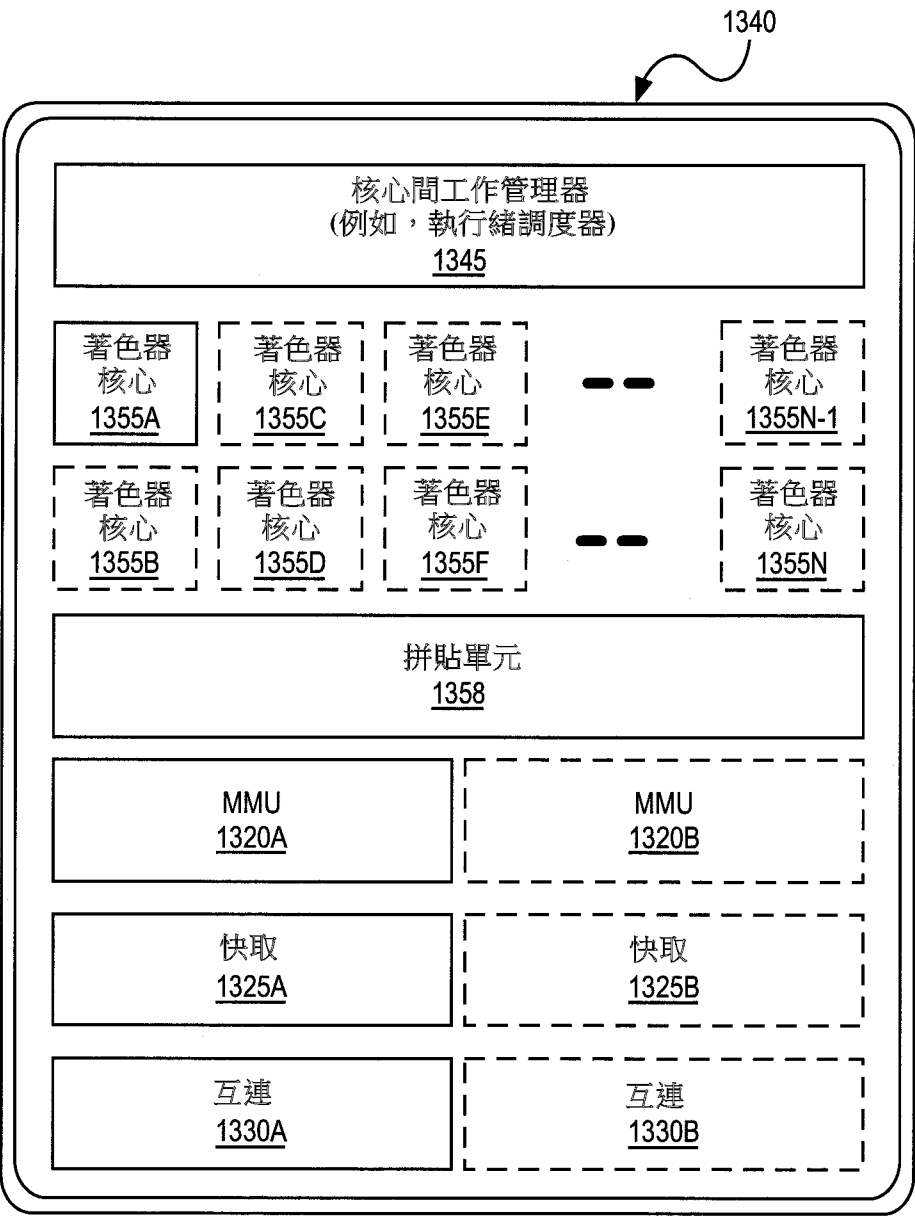
【圖 11D】



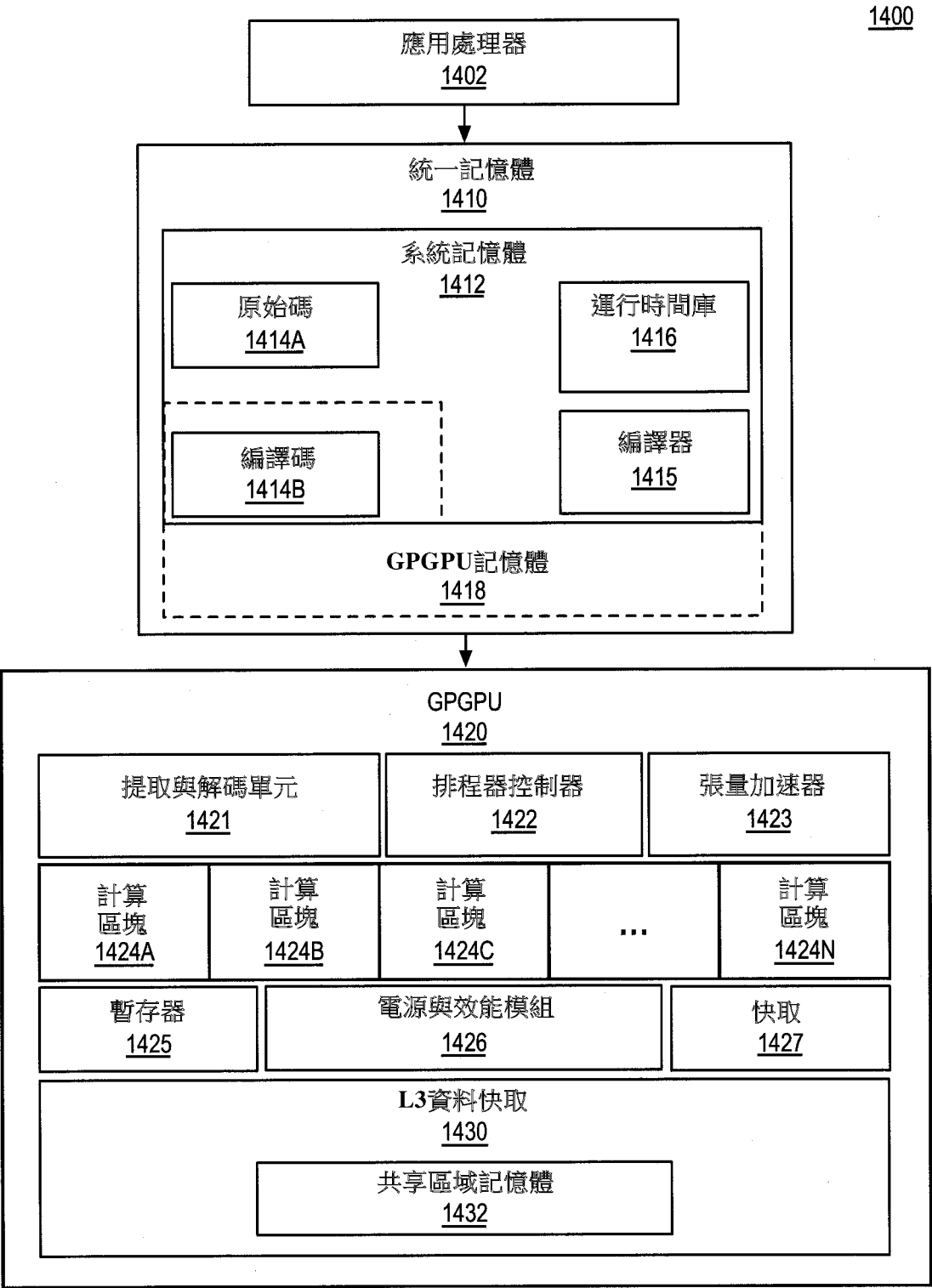
【圖 12】



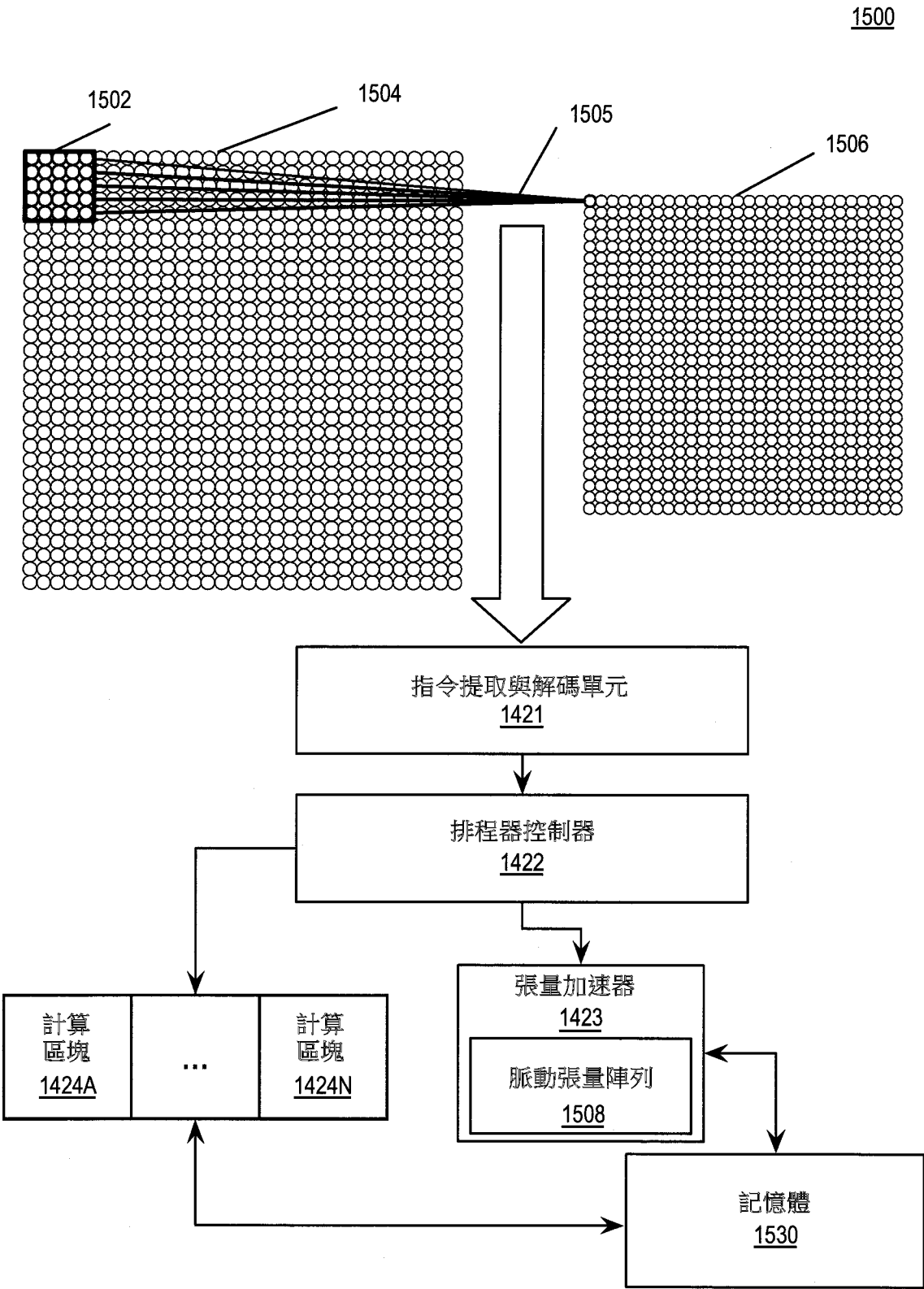
【圖 13A】



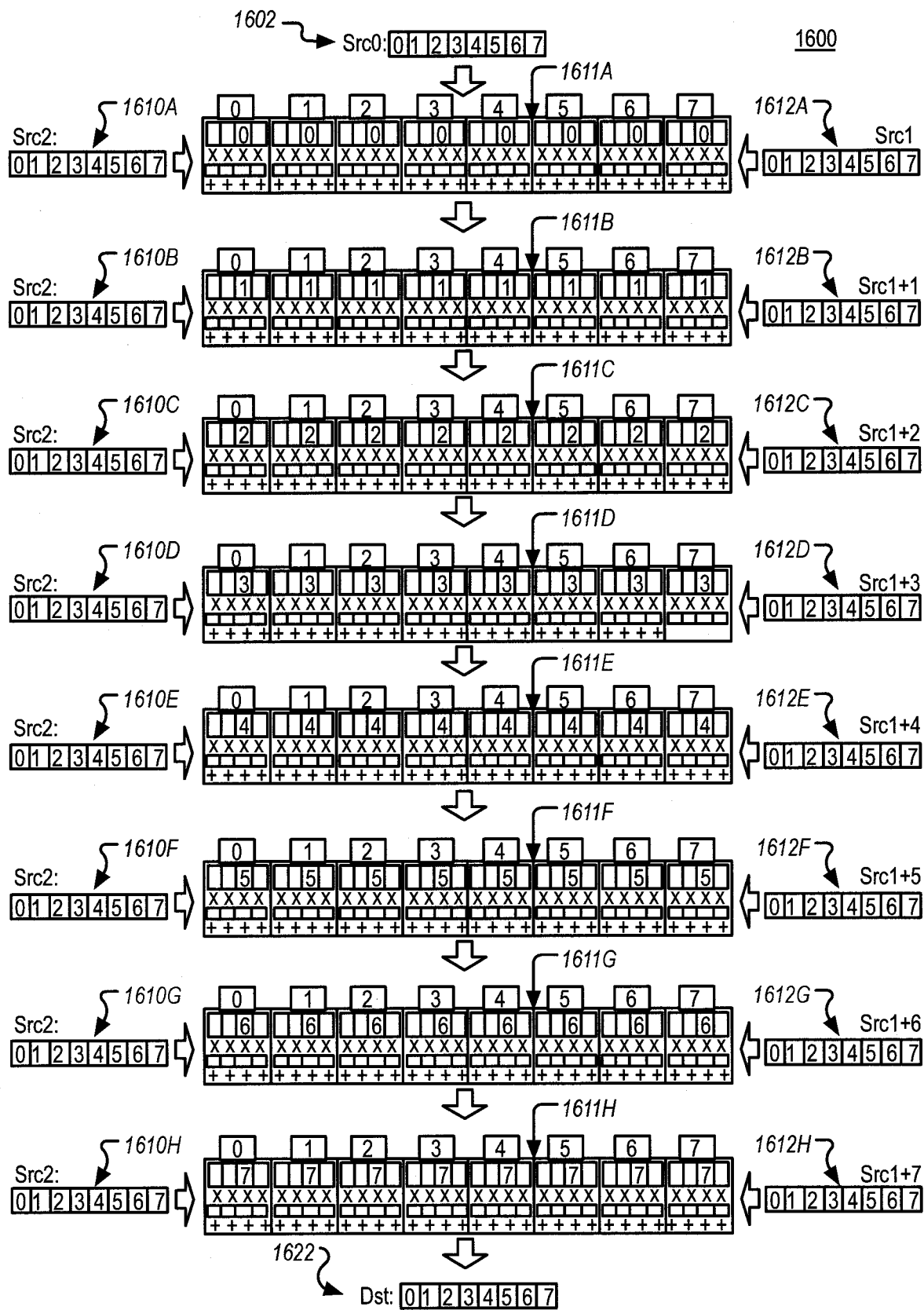
【圖 13B】



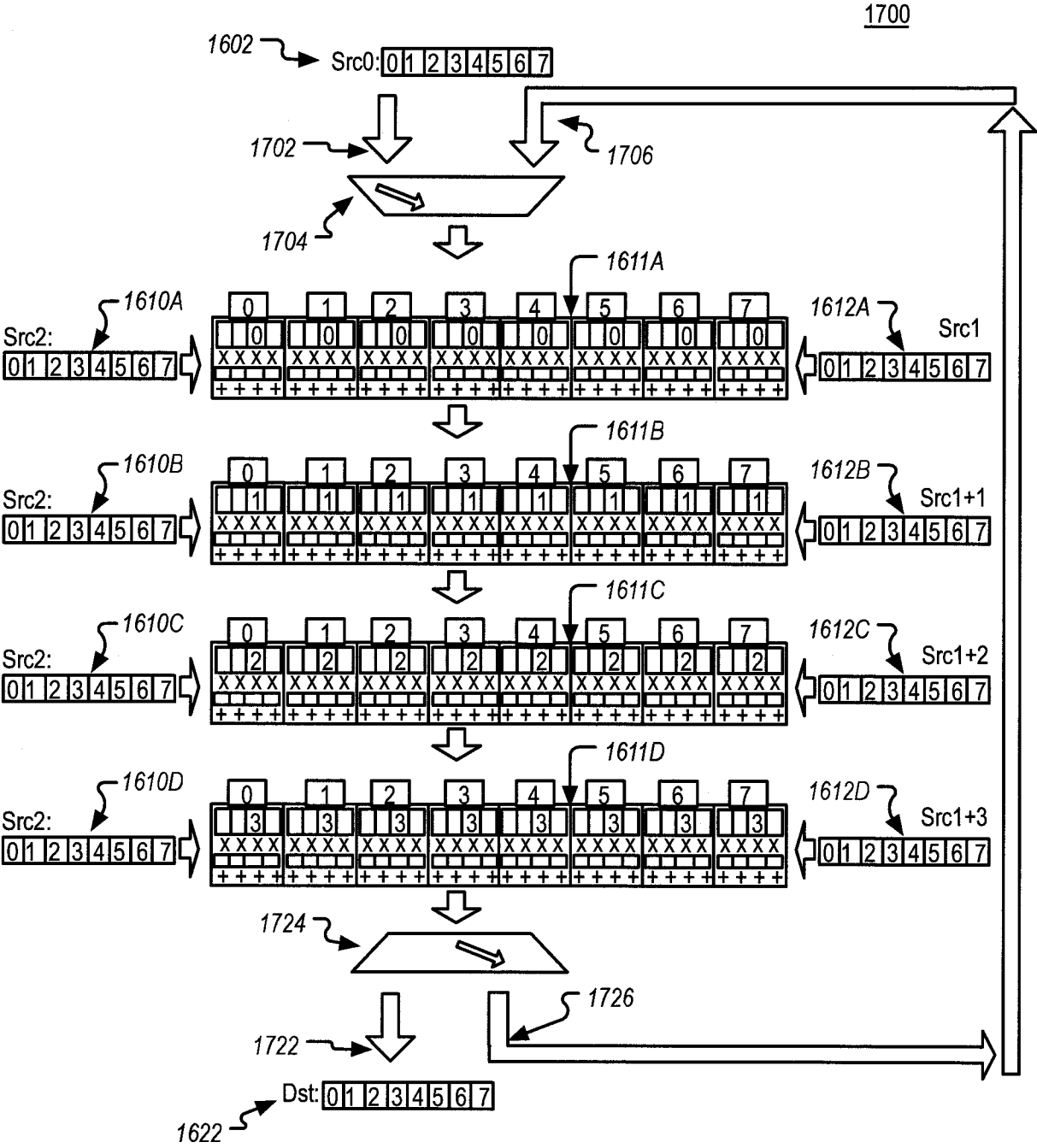
【圖 14】



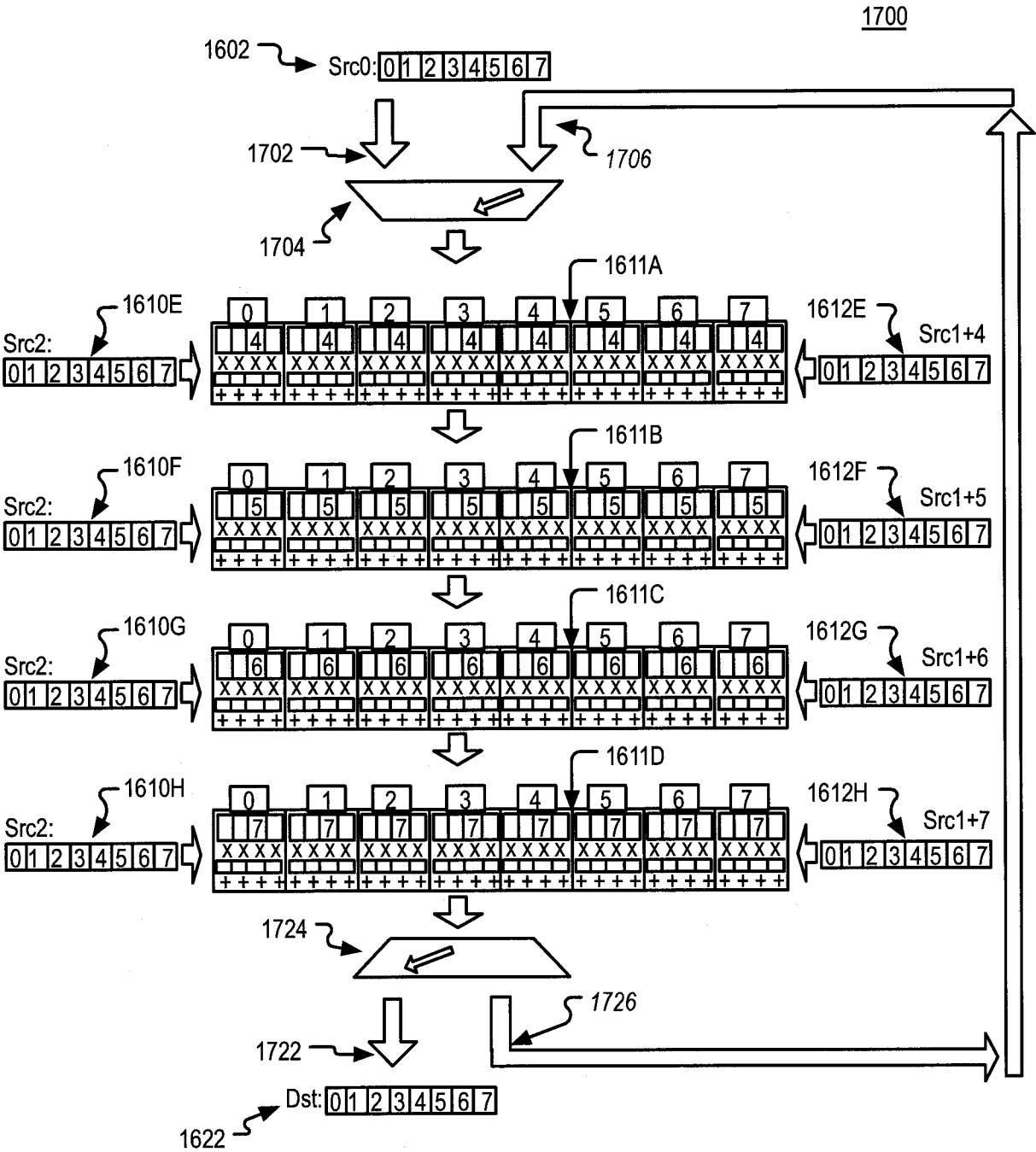
【圖 15】



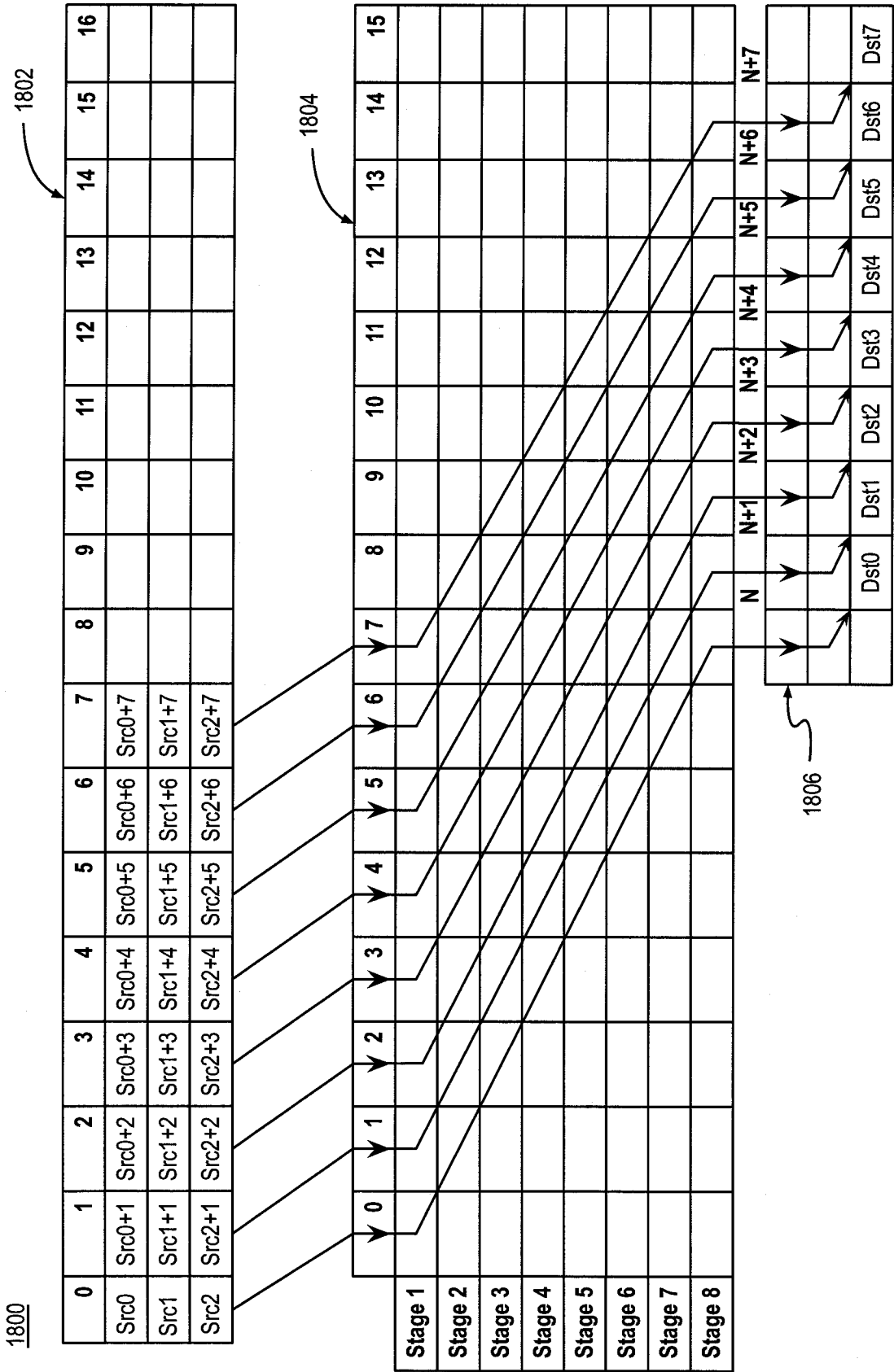
【圖 16】



【圖 17A】



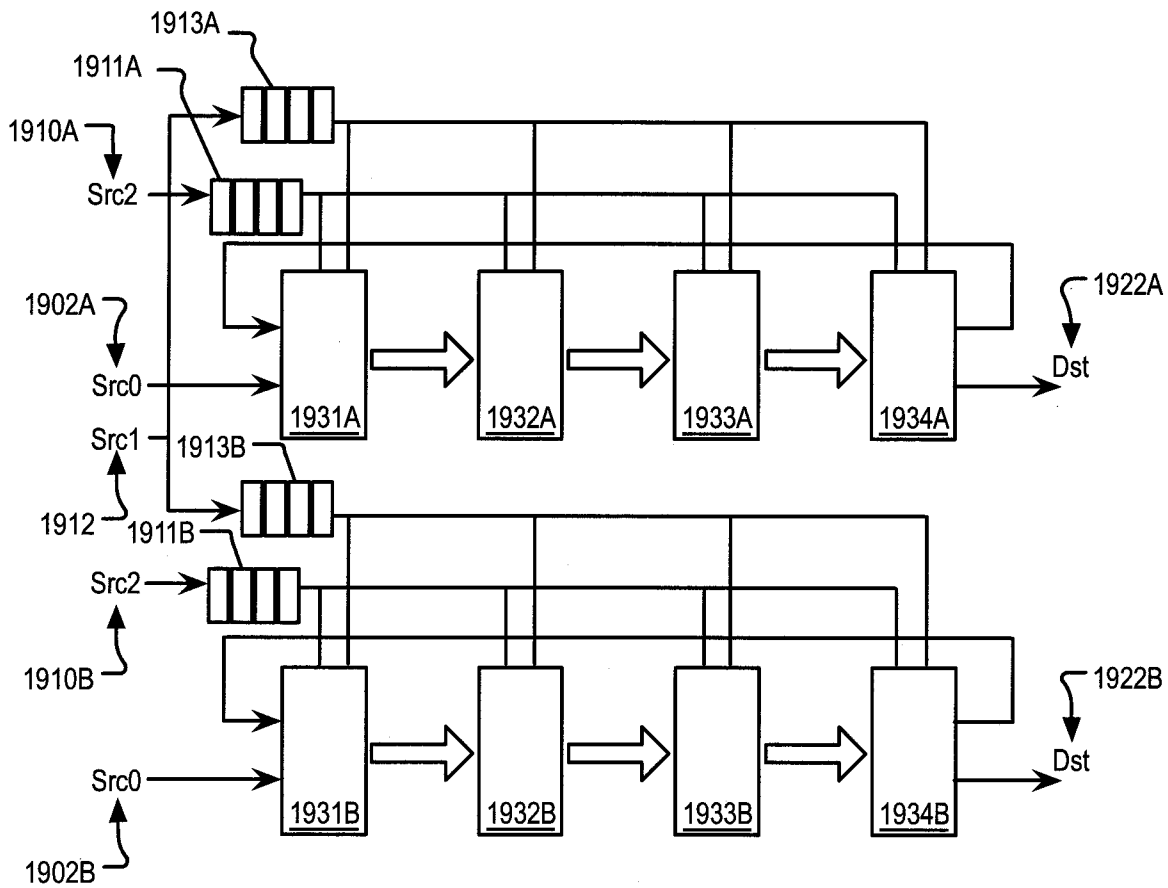
【圖 17B】



【圖 18A】

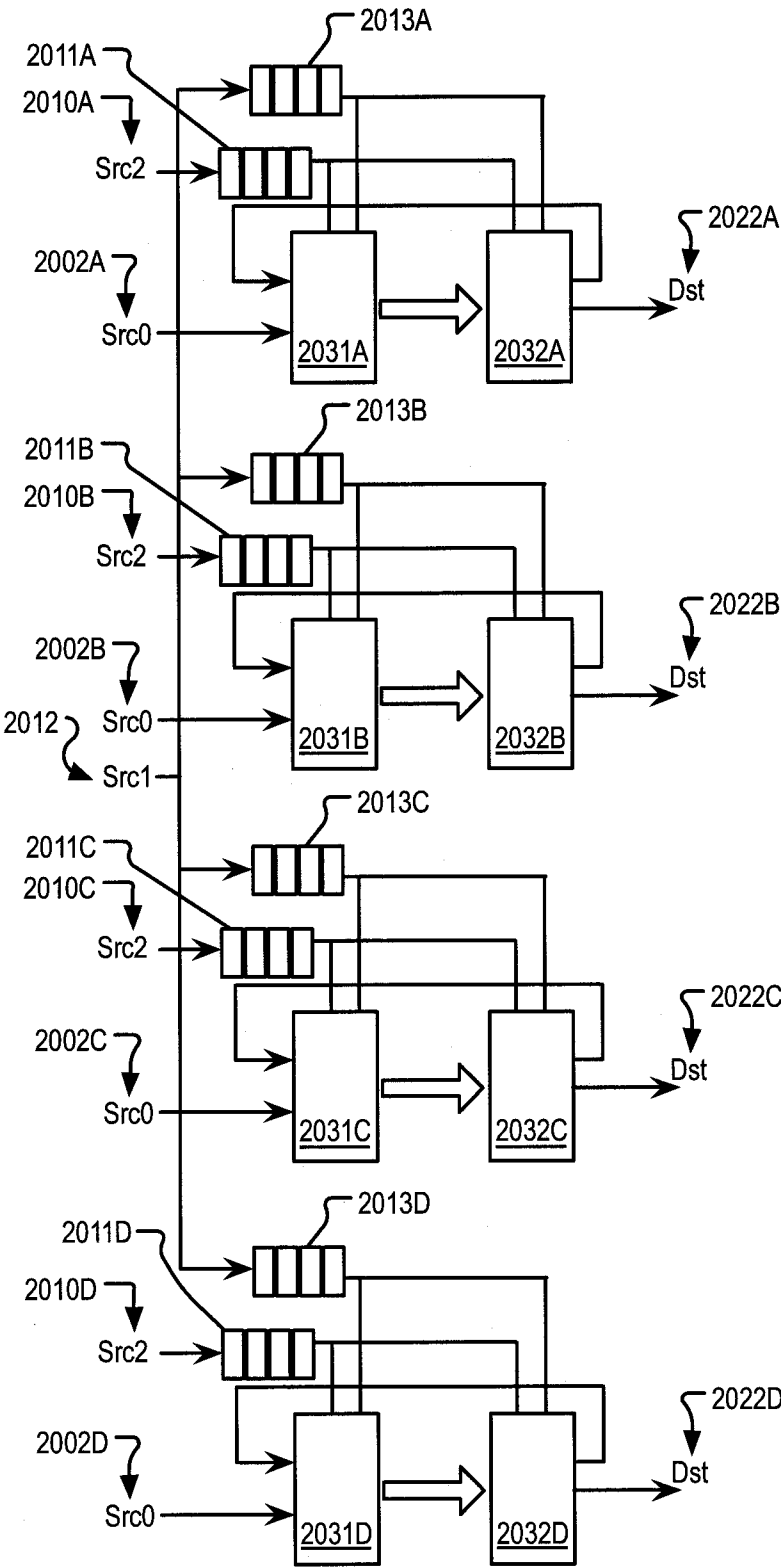


1900

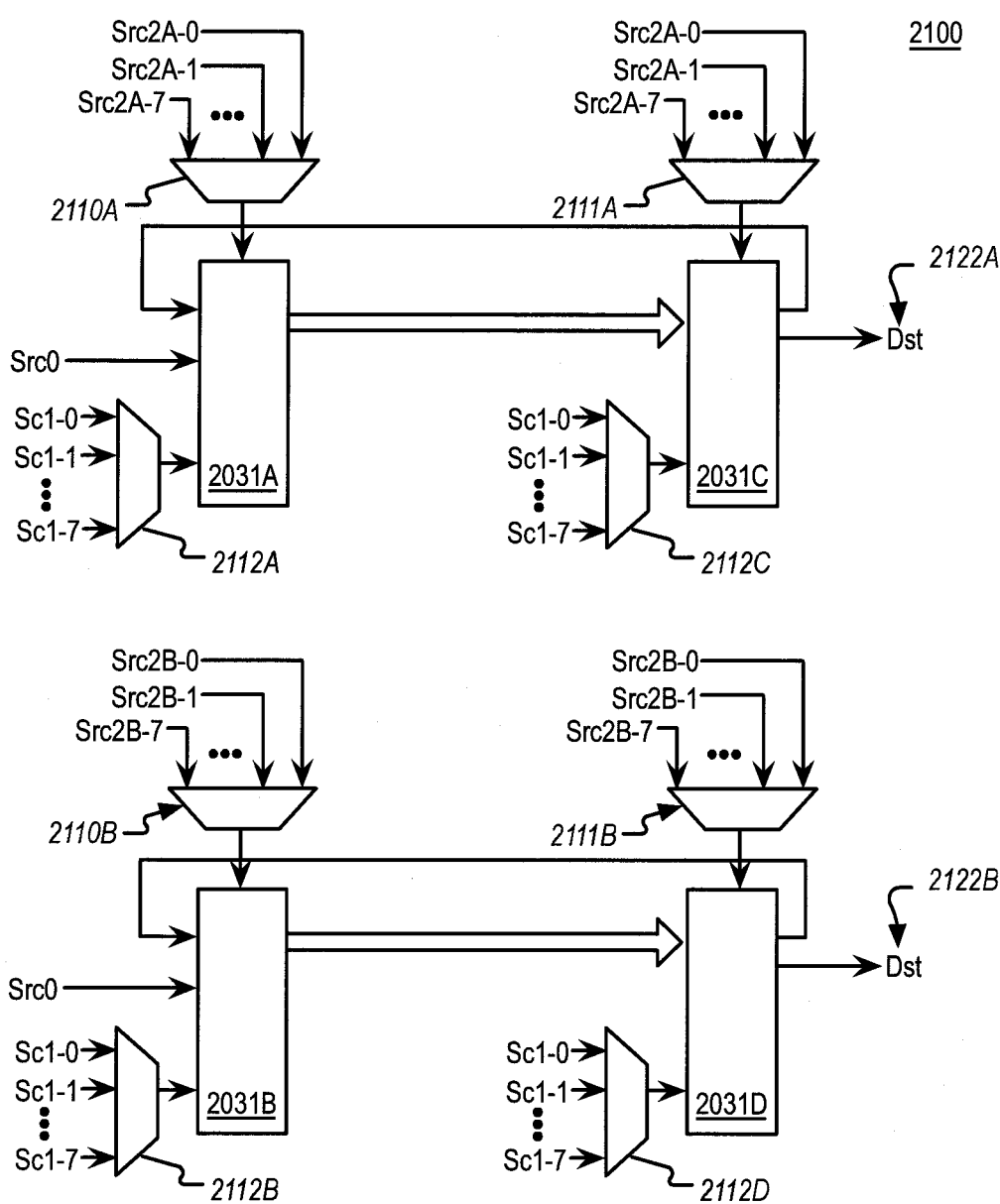


【圖 19】

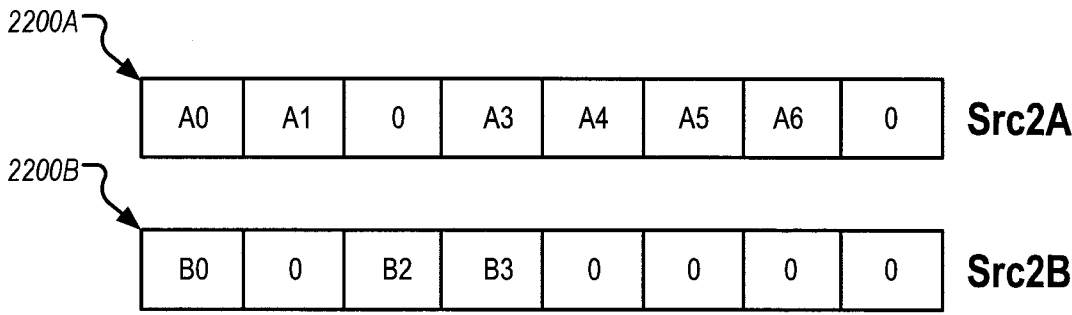
2000



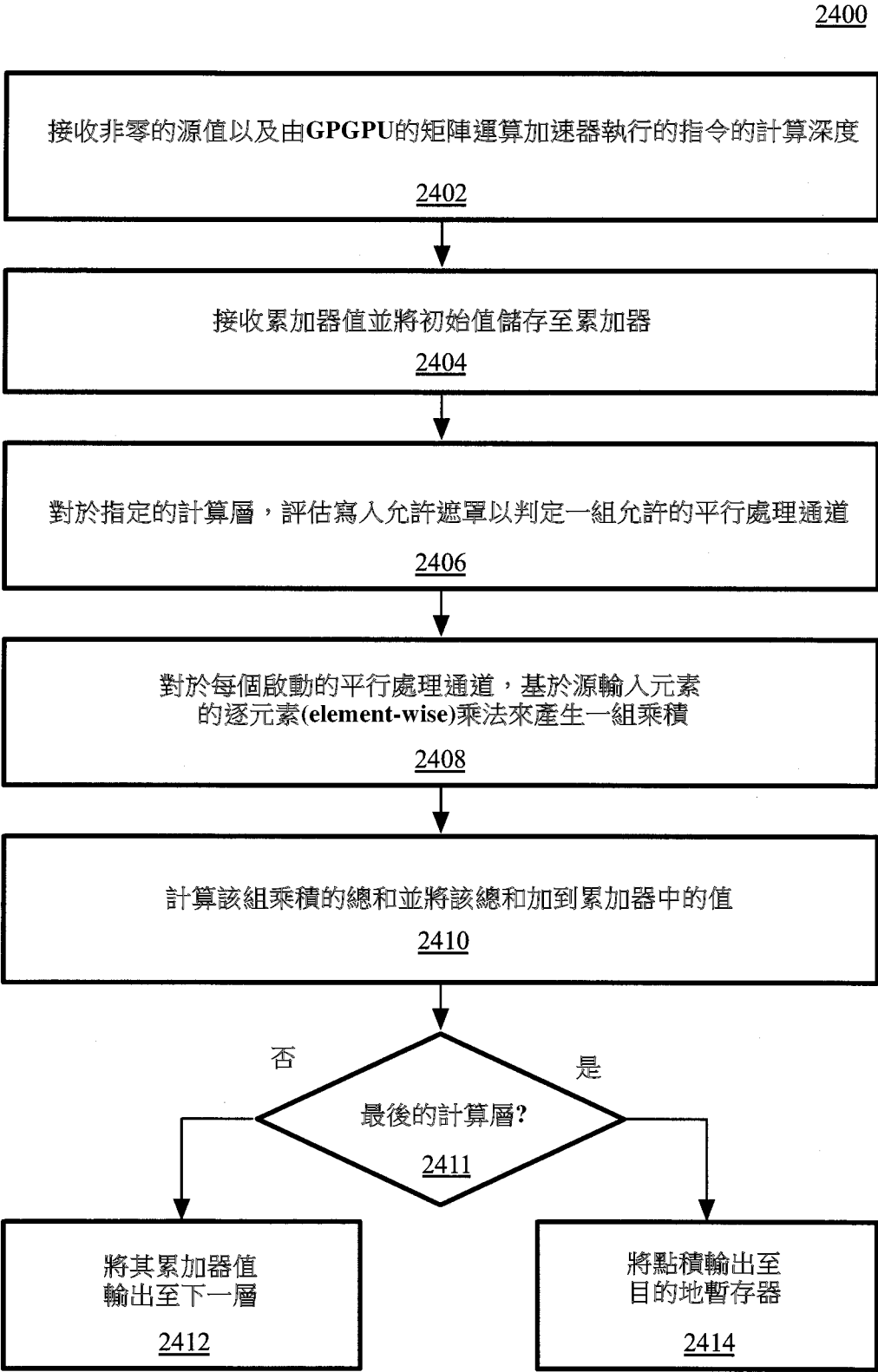
【圖 20】



【圖 21】

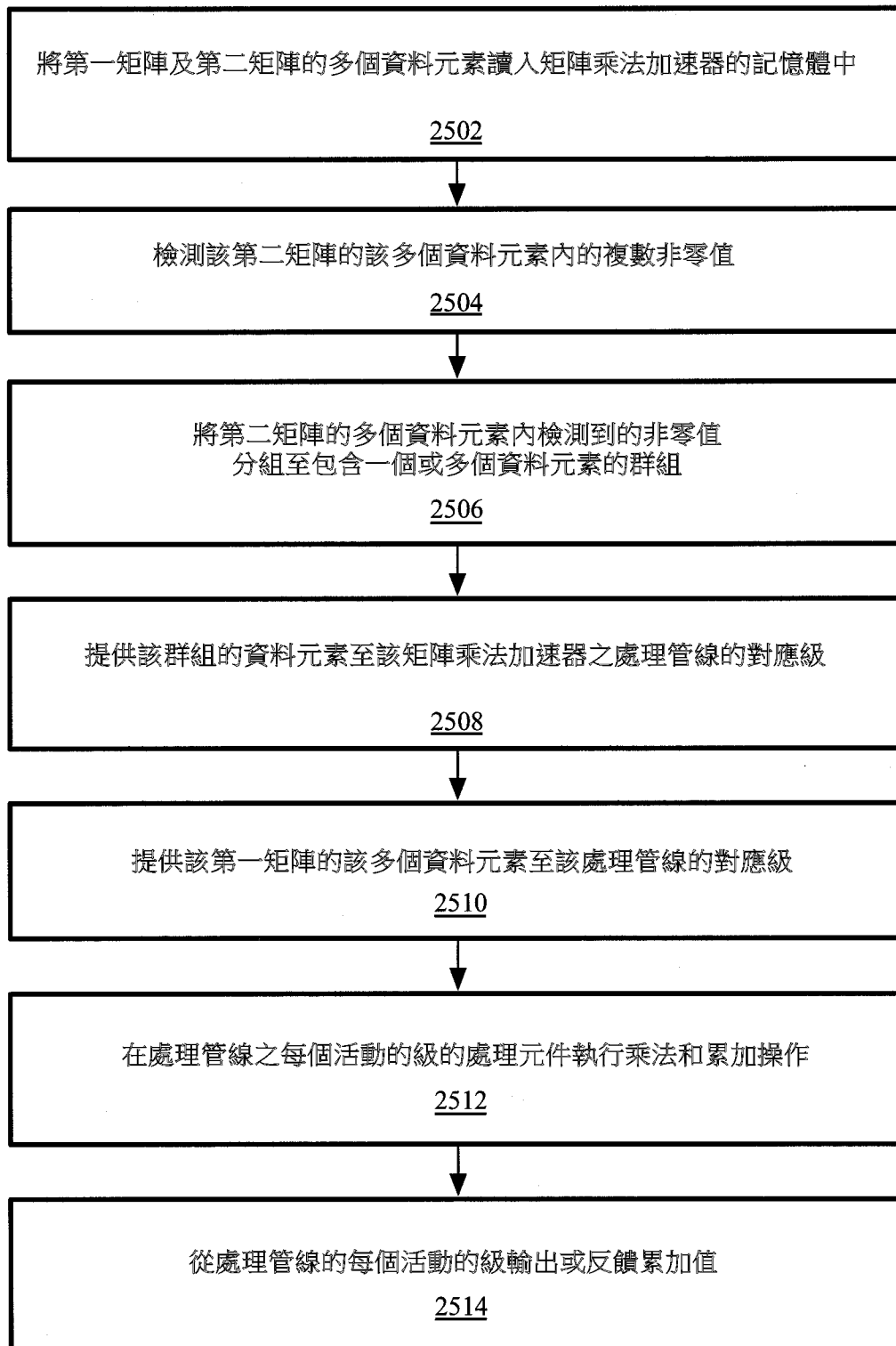


【圖 22】

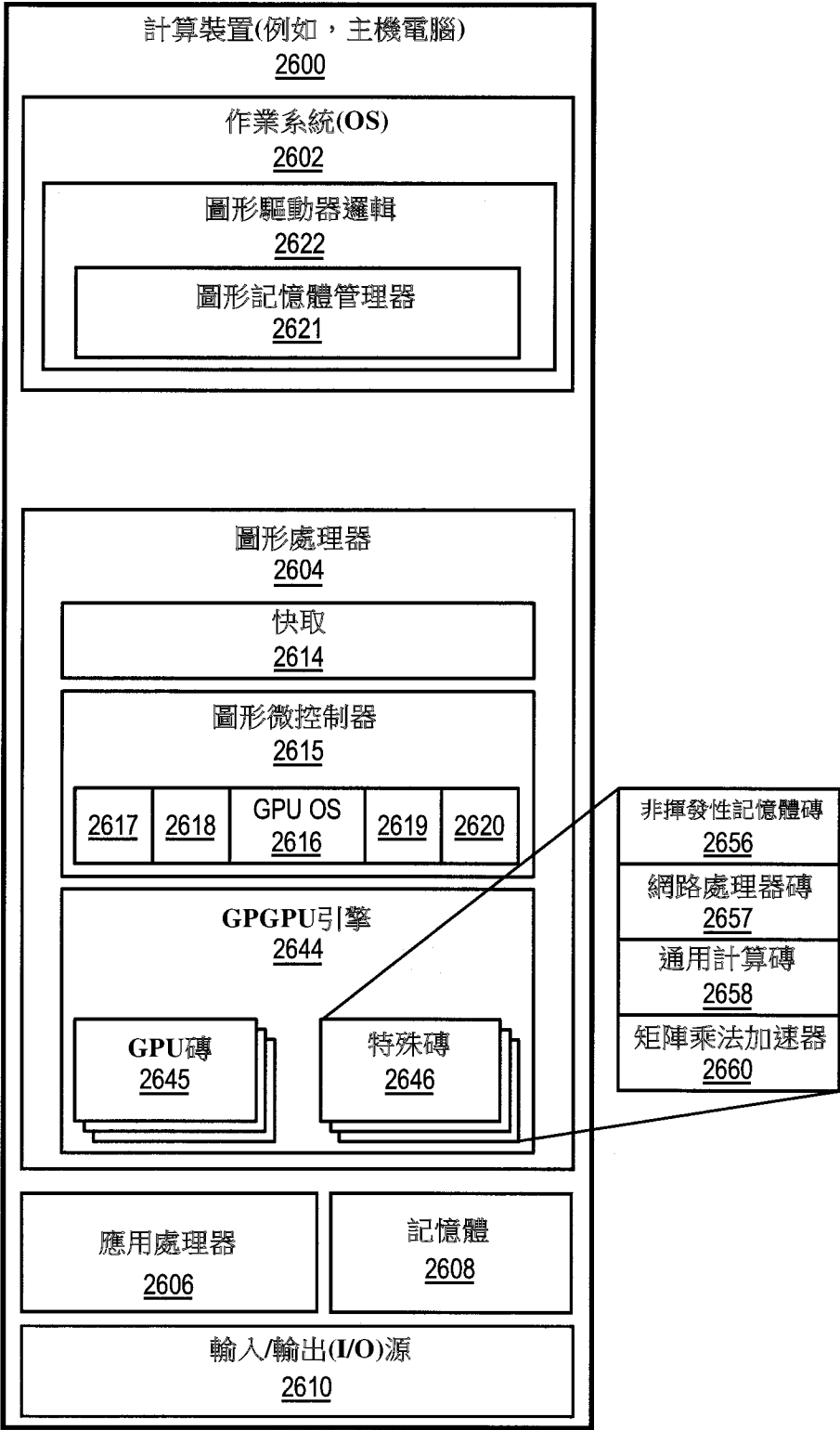


【圖 24】

2500



【圖 25】



【圖 26】