



(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 I846980 B

(45)公告日：中華民國 113 (2024) 年 07 月 01 日

(21)申請案號：109139554 (22)申請日：中華民國 109 (2020) 年 11 月 12 日

(51)Int. Cl. : G06T9/00 (2006.01) G06N3/04 (2023.01)
G06T15/00 (2011.01)

(30)優先權：2019/11/15 美國 62/935,670
2020/11/11 美國 17/095,544

(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：亞布 亞奇雪克 APPU, ABHISHEK (IN)；瑟提 普拉庫馬 SURTI,
PRASOONKUMAR (US)；博依思 吉爾 BOYCE, JILL (US)；麥猶蘭 沙布曼尼
恩 MAIYURAN, SUBRAMANIAM (US)；亞波達卡 麥克 APODACA, MICHAEL
(US)；雷克 亞當 LAKE, ADAM T. (US)；何連 詹姆士 HOLLAND, JAMES
(US)；雷岡納森 華森士 RANGANATHAN, VASANTH (US)；寇克 歐塔
KOKER, ALTUG (US)；徐利同 XU, LIDONG (CN)；卡布拉索斯 尼寇斯
KABURLASOS, NIKOS (US)

(74)代理人：林志剛

(56)參考文獻：

TW	201732556A	TW	201732734A
CN	107689948A	US	2019/0108436A1

審查人員：吳家豪

申請專利範圍項數：24 項 圖式數：41 共 266 頁

(54)名稱

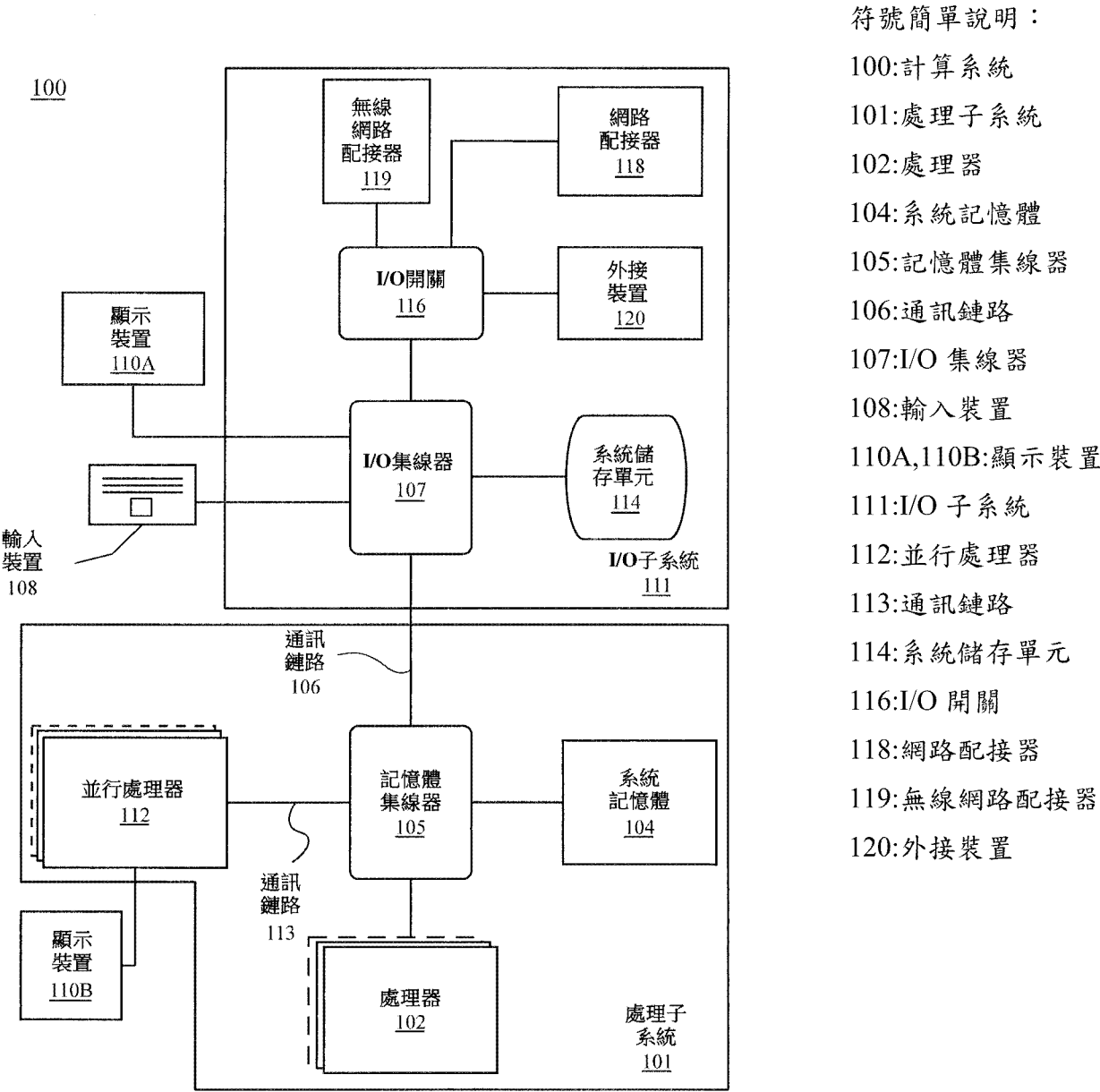
用於稀疏資料之脈動式算術的計算裝置、方法、圖形處理系統、非暫時性機器可讀取媒體及圖形處理器

(57)摘要

文中所述之實施例提供一種指令集相關聯邏輯，以使包括張量加速器之處理資源能夠執行稀疏子矩陣操作的優化計算。一個實施例提供了將數值變換應用至矩陣資料以增加該資料的該稀疏性的硬體邏輯。當壓縮矩陣資料時，增加稀疏性可能導致較高的壓縮率。

Embodiments described herein provided for an instruction and associated logic to enable a processing resource including a tensor accelerator to perform optimized computation of sparse submatrix operations. One embodiment provides hardware logic to apply a numerical transform to matrix data to increase the sparsity of the data. Increasing the sparsity may result in a higher compression ratio when the matrix data is compressed.

指定代表圖：



【圖 1】



I846980

公告本

【發明摘要】

【中文發明名稱】

用於稀疏資料之脈動式算術的計算裝置、方法、圖形處理系統、非暫時性機器可讀取媒體及圖形處理器

【英文發明名稱】

COMPUTING DEVICE, METHOD, GRAPHICS
PROCESSING SYSTEM, NON-TRANSITORY
MACHINE-READABLE MEDIA, AND GRAPHICS
PROCESSOR FOR SYSTOLIC ARITHMETIC ON SPARSE
DATA

【中文】

文中所述之實施例提供一種指令集相關聯邏輯，以使包括張量加速器之處理資源能夠執行稀疏子矩陣操作的優化計算。一個實施例提供了將數值變換應用至矩陣資料以增加該資料的該稀疏性的硬體邏輯。當壓縮矩陣資料時，增加稀疏性可能導致較高的壓縮率。

【英文】

Embodiments described herein provided for an instruction and associated logic to enable a processing resource including a tensor accelerator to perform optimized computation of sparse submatrix operations. One embodiment provides hardware logic to apply a numerical transform to matrix data to increase the sparsity of the data. Increasing the sparsity may result in a higher compression ratio when the matrix data is compressed.

【指定代表圖】第(1)圖。

【代表圖之符號簡單說明】

100:計算系統

101:處理子系統

102:處理器

104:系統記憶體

105:記憶體集線器

106:通訊鏈路

107:I/O集線器

108:輸入裝置

110A,110B:顯示裝置

111:I/O子系統

112:並行處理器

113:通訊鏈路

114:系統儲存單元

116:I/O開關

118:網路配接器

119:無線網路配接器

120:外接裝置

【特徵化學式】無

【發明說明書】

【中文發明名稱】

用於稀疏資料之脈動式算術的計算裝置、方法、圖形處理系統、非暫時性機器可讀取媒體及圖形處理器

【英文發明名稱】

COMPUTING DEVICE, METHOD, GRAPHICS
PROCESSING SYSTEM, NON-TRANSITORY
MACHINE-READABLE MEDIA, AND GRAPHICS
PROCESSOR FOR SYSTOLIC ARITHMETIC ON SPARSE
DATA

【技術領域】

【0001】本揭露大致上係有關於資料處理，並且更具體地是有關於經由通用圖形處理單元的資料處理。

交叉索引

【0002】本申請主張於2019年11月15日申請之美國臨時專利申請案第62/935,670之優先權，特此以引用方式併入本文中。

【先前技術】

【0003】當前的並行圖形資料處理包括被開發以對圖形資料執行諸如(例如)線性內插、鑲嵌、柵格化、紋理映

射、深度測試等特定操作之系統及方法。傳統上，圖形處理器使用固定功能計算單元來處理圖形資料。然而，更近期地，部分的圖形處理器已被製成可編程的，其致使這樣的處理器支持更多樣的操作以用於處理頂點及片段資料。

【0004】為了進一步提高效能，圖形處理器通常實施諸如管線操作處理技術，這些處理技術嘗試在圖形管線的不同部分中並行處理盡可能多的圖形資料。具有單指令多執行緒(SIMT)架構的並行圖形處理器旨在最大化圖形管線中的並行處理量。在SIMT架構中，數組並行執行緒嘗試盡可能經常同步執行程式指令以提高處理效率。針對SIMT架構之軟體及硬體的一般概述可見於Shane Cook, *CUDA Programming*, 第3章，第37-51頁(2013)。

【圖式簡單說明】

【0005】為了能夠詳細理解本實施例的上述特徵的方式，可以通過參考實施例來對上面簡要概述的實施例進行更具體的描述，其中一些實施例在附圖中繪示。然而應注意的是，附圖僅繪示了典型的實施例，因此不應被視為限制其範圍。

【0006】[圖1]係繪示組態以實施文中所述之實施例的一或多個態樣之電腦系統的方塊圖；

【0007】[圖2A-2D]繪示並行處理器組件；

【0008】[圖3A-3C]係根據實施例之圖形多處理器的方塊圖；

【0009】[圖 4A-4F]繪示其中複數個 GPU 被通訊地耦接至複數個多核心處理器的例示性架構；

【0010】[圖 5]繪示圖形處理管線；

【0011】[圖 6]繪示機器學習軟體堆疊；

【0012】[圖 7]繪示通用圖形處理單元；

【0013】[圖 8]繪示多 GPU 計算系統；

【0014】[圖 9A-9B]繪示例示性深度類神經網路模型之層；

【0015】[圖 10]繪示例示性遞迴類神經網路；

【0016】[圖 11]繪示深度類神經網路之訓練和部署。

【0017】[圖 12A]係繪示分散式學習的方塊圖；

【0018】[圖 12B]係繪示用以加速分散式學習之可編程網路介面的方塊圖；

【0019】[圖 13]繪示適用於使用經訓練的模型來執行推理的例示性推理系統單晶片(SOC)；

【0020】[圖 14]係繪示微影系統的方塊圖。

【0021】[圖 15A-15C]繪示計算系統及圖形處理器；

【0022】[圖 16A-16C]繪示額外的圖形處理器和計算加速器架構的方塊圖；

【0023】[圖 17]係圖形處理器之圖形處理引擎的方塊圖；

【0024】[圖 18A-18B]繪示包括在圖形處理器核心中採用之處理元件陣列的執行緒執行邏輯；

【0025】[圖 19]繪示額外的執行單元；

【0026】[圖 20]係繪示圖形處理器指令格式的方塊圖；

【0027】[圖 21]係額外的圖形處理器架構的方塊圖；

【0028】[圖 22A-22B]繪示圖形處理器命令格式及命令序列；

【0029】[圖 23]繪示用於資料處理系統之例示性圖形軟體架構；

【0030】[圖 24A]係繪示IP核心開發系統的方塊圖；

【0031】[圖 24B]繪示積體電路封裝總成的橫斷面側視圖；

【0032】[圖 24C]繪示包括連接至基材(例如，基底晶粒)之多個硬體邏輯小晶片單元的封裝總成；

【0033】[圖 24D]繪示包括可互換小晶片的封裝總成；

【0034】[圖 25]係繪示例示性系統單晶片積體電路的方塊圖；

【0035】[圖 26A-26B]係繪示在SoC內使用的例示性圖形處理器的方塊圖；

【0036】[圖 27]繪示根據一實施例之額外的執行單元；

【0037】[圖 28]繪示根據實施例之由指令管線執行的矩陣操作；

【0038】[圖 29A-29B]繪示根據一些實施例之基於硬體的脈動張量陣列的細節；

【0039】[圖 30]繪示根據實施例之包括具有稀疏優化的脈動張量陣列的處理資源；

【0040】[圖 31A-31B]繪示根據實施例之旁路零值子矩陣的系統；

【0041】[圖 32]繪示計算資源，其包括繞過脈動張量陣列以用於在子矩陣上操作的邏輯，該子矩陣包括單一非零值；

【0042】[圖 33]繪示根據一實施例之組態以使得能夠壓縮傳輸類神經網路資料的計算架構；

【0043】[圖 34A-34B]繪示根據實施例之用於稀疏類神經網路資料的顯著圖編碼；

【0044】[圖 35]繪示使用顯著圖資料來促進對脈動張量陣列的元件的旁路或時脈/功率門控；

【0045】[圖 36]繪示根據一實施例之將類神經網路資料轉換為緊湊的非零係數的系統；

【0046】[圖 37A-37B]繪示對稀疏類神經網路資料執行變換和逆變換的邏輯單元；

【0047】[圖 38A-38B]繪示在圖形處理器上產生和使用變換後的矩陣的方法；

【0048】[圖 39]繪示被組態以對壓縮或編碼的資料進行操作的脈動張量陣列的版本；以及

【0049】[圖 40]繪示對壓縮或編碼的資料執行脈衝式算術的方法；以及

【0050】[圖 41]係根據實施例之包括圖形處理器之計

算裝置的方塊圖。

【發明內容】及【實施方式】

【0051】圖形處理單元(GPU)係通訊地耦接至主機/處理器核心以加速(例如)圖形操作、機器學習操作、模式分析操作及/或各種通用GPU(GPGPU)功能。GPU可透過匯流排或另一互連(例如，諸如PCIe或NVLink等高速互連)被通訊地耦接至主機處理器/核心。替代地，GPU可被集成於如該等核心之相同封裝或晶片上，且透過內部處理器匯流排/互連(亦即，封裝或晶片內部)而被通訊地耦接至該些核心。無論GPU被連接的方式為何，處理器核心可依工作描述符所含之命令/指令的序列之形式將工作分配給GPU。GPU接著使用專用電路/邏輯以便有效率地處理這些命令/指令。

【0052】在下面的敘述中，闡述了許多具體細節，以提供更透徹的理解。然而，對於本領域技術人員顯而易見的是，本發明可以在沒有一或多個這些具體細節的情況下實施。在其他情況下，未敘述眾所周知的特徵，以避免模糊本實施例的細節。

系統概述

【0053】圖1係繪示組態成實現本文中所述之實施例的一或多種態樣之計算系統100的方塊圖。計算系統100包括處理子系統101，其具有一或多個處理器102及系統記憶

體 104，其經由可包括記憶體集線器 105 之互連路徑來通訊。記憶體集線器 105 可以是晶片組組件內的獨立組件或可被整合於一或多個處理器 102 內。記憶體集線器 105 係經由通訊鏈路 106 而與 I/O 子系統 111 耦接。I/O 子系統 111 包括 I/O 集線器 107，其可致使計算系統 100 用以接收來自一或多個輸入裝置 108 之輸入。此外，I/O 集線器 107 可致使顯示控制器(其可被包括在一或多個處理器 102 中)用以提供輸出至一或多個顯示裝置 110A。在一個實施例中，與 I/O 集線器 107 耦接中的一或多個顯示裝置 110A 可包括本地的、內部的或嵌入式顯示裝置。

【0054】處理子系統 101(例如)包括經由匯流排或其他通訊鏈路 113 而耦接至記憶體集線器 105 的一或多個並行處理器 112。通訊鏈路 113 可以是任何數目之標準為基的通訊鏈路技術或協定之一，諸如(但不限於)PCI Express；或可以是供應商特定的通訊介面或通訊結構。所述一或多個並行處理器 112 形成一種計算集中的並行或向量處理系統，其可包括大量的處理核心及/或處理叢集，諸如許多積體核心(MIC)處理器。例如，所述一或多個並行處理器 112 形成圖形處理子系統，其可向經由 I/O 集線器 107 耦接至一或多個顯示裝置 110A 中的一者輸出像素。所述一或多個並行處理器 112 亦可包括顯示控制器和顯示介面(未顯示)，以實現直接連接至一或多個顯示裝置 110B。

【0055】在 I/O 子系統 111 內，系統儲存單元 114 可連接至 I/O 集線器 107，以提供用於計算系統 100 之儲存機

制。I/O開關 116可被用來提供介面機制，以致使I/O集線器 107與其他組件之間的連接，諸如可被整合在平台中的網路配接器 118及/或無線網路配接器 119，以及可經由一或多個外接裝置 120而被加入之各種其他裝置。外接裝置 120可包括(例如)一或多個外部圖形處理器裝置、圖形卡、及/或計算加速器。網路配接器 118可為乙太網路配接器或另一有線網路配接器。無線網路配接器 119可包括 Wi-Fi、藍芽、近場通訊(NFC)、或包括一或多個無線電之其他網路裝置中的一或多個。

【0056】計算系統 100可包括未明確繪示的其他組件(包括USB或其他埠連接、光學儲存器驅動、視訊擷取裝置等)，其亦可連接到I/O集線器 107。與圖 1中各組件互連的通訊路徑可使用任何合適的協定實施，所述協定諸如基於協定(例如，PCI-Express)的PCI(週邊組件互連)或任何其他匯流排或點對點通訊介面及/或協定，諸如NVLink高速互連、計算高速鏈路™(CXL™)(例如CXL.mem)、Infinity Fabric (IF)、乙太網路(IEEE 802.3)、遠端直接記憶體存取(RDMA)、InfiniBand、網際網路廣域 RDMA 協定(iWARP)、傳輸控制協定(TCP)、使用者資料報協定(UDP)、快速UDP網際網路連接(QUIC)、基於聚合乙太網路的RDMA(RoCE)、英特爾快速路徑互連(QPI)、英特爾超路徑互連(UPI)、英特爾晶載系統結構(IOSF)、OmniPath、超傳輸、先進微控制器匯流排架構(AMBA)互連、OpenCAPI、Gen-Z、加速器的快取一致性互連(CCIX)、

3GPP長期演進(LTE)(4G)、3GPP 5G及其變化、或本領域已知的有線或無線互連協定。在一些實例中，可使用諸如非揮發性記憶體儲存標準(NVMe)跨結構(NVMe-oF)或NVMe的協定而將資料複製或儲存至虛擬儲存節點

【0057】所述一或多個並行處理器112結合針對圖形和視訊處理而優化的電路，包括(例如)視訊輸出電路，並且組成圖形處理單元(GPU)。替代地或額外地，所述一或多個並行處理器112結合了針對通用處理而被優化的電路，同時保留基本計算架構，更詳細地描述於本文中。計算系統100之組件可與一或多個其他系統元件被整合在單一積體電路上。例如，一或多個並行處理器112、記憶體集線器105、處理器102及I/O集線器107可被整合在系統單晶片(SoC)積體電路之中。替代地，計算系統100之組件可被整合在單一封裝之中，以形成系統級封裝(SIP)組態。在一個實施例中，計算系統100的組件之至少一部分可被整合在多晶片模組(MCM)之中，其可與其他多晶片模組被互連入模組式計算系統中。

【0058】應理解到，本文中所示之計算系統100為說明性的且其變化及修改是可能的。連接拓撲(包括橋的數目和配置、處理器102的數目及並行處理器112的數目)可如所需地修改。例如，系統記憶體104可直接地連接至處理器102而不是透過橋，而其他裝置係經由記憶體集線器105及處理器102與系統記憶體104通訊。在其他替代拓撲中，並行處理器112被連接至I/O集線器107或直接至一或

多個處理器 102 之一，而不是至記憶體集線器 105。在其他實施例中，I/O 集線器 107 和記憶體集線器 105 可被整合在單一晶片之中。還可經由多個插口而相連之兩組或更多組的處理器 102，其可與並行處理器 112 中的兩或更多實例耦接。

【0059】本文中所示之一些特定組件為選擇性的且可能不被包括在計算系統 100 的所有實現中。例如，任何數目的外接卡或週邊設備可被支持，或者一些組件可被刪除。再者，一些架構可對於類似圖 1 中所示之那些組件使用不同術語。例如，記憶體集線器 105 可在一些架構中被稱為北橋，而 I/O 集線器 107 可被稱為南橋。

【0060】圖 2A 繪示並行處理器 200。所述並行處理器 200 可為 GPU、GPGPU、或本文所述的相似物。並行處理器 200 的各個組件可使用一或多個積體電路裝置來實現，諸如可編程處理器、特殊應用積體電路 (ASIC) 或場可編程閘陣列 (FPGA)。所繪示的並行處理器 200 可為圖 1 中所示之一或多個處理器 112。

【0061】所述並行處理器 200 包括並行處理單元 202。並行處理單元包括致使與其他裝置通訊之 I/O 單元 204，其包括並行處理單元 202 的其他實例。I/O 單元 204 可直接連接至其他裝置。例如，I/O 單元 204 經由集線器或開關介面 (諸如，記憶體集線器 105) 的使用與其他裝置連接。介於記憶體集線器 105 與 I/O 單元 204 之間的連接形成通訊鏈路 113。在並行處理單元 202 之內，I/O 單元 204 係與主介面

206及記憶體交叉開關216連接，其中主介面206接收了用於執行處理操作的命令，而記憶體交叉開關216接收了用於執行記憶體操作的命令。

【0062】當主介面206經由I/O單元204接收了命令緩衝器時，主介面206可將用以執行那些命令之工作操作指引至前端208。在一實施例中，前端208與排程器210耦接，排程器210係組態以將命令或其他工作項目分配至處理叢集陣列212。在工作被分配至處理叢集陣列212之處理叢集之前，排程器210確保處理叢集陣列212被適當地組態且處於有效狀態。排程器210經由在微控制器上執行的軟體邏輯來實現。微控制器實現的排程器210可配置以在粗略和精細粒度執行複雜的排程和工作分配的操作，致使快速先佔和執行緒的背景切換執行在處理叢集陣列212上。較佳地，主機軟體可以證明工作負載經由多個圖形處理門鈴中的一個來排程在處理叢集陣列212上。在其他實例中，對新工作負載或中斷的輪詢可用於標識或指示要執行的工作的可用性。工作負載接著可以藉由排程器微控制器內的排程器210邏輯來自動地跨處理叢集陣列212分佈。

【0063】處理叢集陣列212可包括高達「N」個處理叢集(例如，叢集214A、叢集214B至叢集214N)。處理叢集陣列212之每一個叢集214A-214N可執行大量的並行執行緒。排程器210可使用各種排程及/或工作分配演算法以將工作配置給處理叢集陣列212中的叢集214A-214N，其可根據各類型的程式或計算所產生的工作負載而改變。排程

可由排程器 210 來動態地處置，或在針對藉由處理叢集陣列 212 之執行而組態的程式邏輯之編譯期間可部分地由編譯器邏輯來協助。可選地，處理叢集陣列 212 的不同叢集 214A-214N 可被配置以供處理不同類型的程式或以供執行不同類型的計算。

【0064】處理叢集陣列 212 可被組態以執行各種類型的並行處理操作。在一實施例中，處理叢集陣列 212 被組態以執行通用並行計算操作。例如，處理叢集陣列 212 可包括用以執行處理工作之邏輯，該些處理工作包括視訊及/或音頻資料之過濾、執行包括物理操作的建模操作及執行資料轉換。

【0065】處理叢集陣列 212 被組態為執行並行圖形處理操作。在其中並行處理器 200 被組態為執行圖形處理操作的實施例中，處理叢集陣列 212 可包括支持這些圖形處理操作之執行的額外邏輯，包括但不限於執行紋理操作的紋理採樣邏輯，以及鑲嵌邏輯和其他頂點處理邏輯。此外，處理叢集陣列 212 可以被組態為執行與圖形處理相關的著色器程式，諸如但不限於頂點著色器、鑲嵌著色器、幾何著色器和像素著色器。並行處理單元 202 可以經由 I/O 單元 204 將來自系統記憶體之資料轉移以供處理。在處理期間，已轉移資料可於處理期間被儲存至晶載記憶體(例如，並行處理器記憶體 222)，接著被寫回至系統記憶體。

【0066】在其中並行處理單元 202 係用來執行圖形處理的實施例中，排程器 210 可經組態以將處理工作負載劃

分為約略相等大小的工作，以較佳地致使圖形處理操作之分配至處理叢集陣列 212 中的多個叢集 214A-214N。在一些實施例中，部分的處理叢集陣列 212 可被組態以執行不同類型的處理。例如，第一部分可被組態成執行頂點著色及拓撲產生，第二部分可被組態成執行鑲嵌及幾何著色，而第三部分可被組態成執行像素著色或其他螢幕空白操作，以產生渲染影像以供顯示。由叢集 214A-214N 中的一或多者所產生的中間資料可被儲存於緩衝器中，以容許該中間資料被傳輸在叢集 214A-214N 之間以供進一步處理。

【0067】在操作期間，處理叢集陣列 212 可經由排程器 210 來接收待執行的處理工作，其接收來自前端 208 之界定處理工作的命令。針對處理操作，處理工作可包括待處理資料之指標，例如，表面(修補)資料、基元資料、頂點資料及/或像素資料、以及界定該資料應如何被處理(例如，哪個程式應被執行)之狀態參數和命令。排程器 210 可被組態成提取相應於該些工作之指標或者可接收來自前端 208 之指標。前端 208 可被組態成在由到來的命令緩衝器(例如，批次緩衝器、推送緩衝器，等等)所指明之工作負載被起始之前，確保處理叢集陣列 212 被組態為有效狀態。

【0068】並行處理單元 202 中的一或多個實例之各者可與並行處理器記憶體 222 耦接。並行處理器記憶體 222 可經由記憶體交叉開關 216 而被存取，其可接收來自處理叢集陣列 212 以及 I/O 單元 204 之記憶體請求。記憶體交叉開

關 216 可經由記憶體介面 218 來存取並行處理器記憶體 222。記憶體介面 218 可包括多個分割單元(例如，分割單元 220A、分割單元 220B 至分割單元 220N)，其可各自耦接至並行處理器記憶體 222 的一部分(例如，記憶體單元)。分割單元 220A-220N 的數目被組態成等於記憶體單元的數目，以致使第一分割單元 220A 具有相應的第一記憶體單元 224A、第二分割單元 220B 具有相應的第二記憶體單元 224B 及第 N 分割單元 220N 具有相應的第 N 記憶體單元 224N。在其他實施例中，分割單元 220A-220N 的數目可能不等於記憶體裝置的數目。

【0069】記憶體單元 224A-224N 可包括各種類型的記憶體裝置，包括動態隨機存取記憶體(DRAM)或圖形隨機存取記憶體，諸如同步圖形隨機存取記憶體(SGRAM)，包括圖形雙資料速率(GDDR)記憶體。可選地，記憶體單元 224A-224N 還可包括 3D 堆疊記憶體，包括但不限於高頻寬記憶體(HBM)。本領域技術人員將理解，記憶體單元 224A-224N 之特定實現可變化，並可被選自各種傳統設計之一。渲染目標(諸如框緩衝器或紋理映射)可跨越記憶體單元 224A-224N 被儲存，其容許分割單元 220A-220N 並行地寫入各渲染目標之部分，以有效率地使用並行處理器記憶體 222 之可用頻寬。在一些實施例中，並行處理器記憶體 222 的本地實例可被排除以利於統一的記憶體設計，其利用結合本地快取記憶體的系統記憶體。

【0070】可選地，處理叢集陣列 212 的叢集 214A-

214N中的任一者可處理將被寫入至並行處理器記憶體222內的記憶體單元224A-224N中的任一者的資料。記憶體交叉開關216可被組態成將各叢集214A-214N之輸出轉移至任何分割單元220A-220N或者至可對該輸出執行額外的處理操作之另一叢集214A-214N。每一叢集214A-214N可透過記憶體交叉開關216而與記憶體介面218通訊，以從各種外部記憶體裝置讀取或寫入至各種外部記憶體裝置。在具有記憶體交叉開關216之實施例一者中，記憶體交叉開關216具有至記憶體介面218用以與I/O單元204通訊之連接，以及至並行處理器記憶體222之本地實例的連接，其致使不同處理叢集214A-214N內的處理單元與系統記憶體或其他不在並行處理單元202本地之記憶體通訊。大致上，記憶體交叉開關216可使用虛擬通道來分離介於叢集214A-214N與分割單元220A-220N之間的流量串。

【0071】雖然並行處理單元202的單一實例被繪示在並行處理器200內，但並行處理單元202之任何數目的實例均可被包括。例如，並行處理單元202的多個實例可被設置在單一外接卡上，或者多個外接卡可被互連。例如，並行處理器200可為外接裝置(諸如，圖1之外接裝置120)，其可為圖形卡，諸如包括一或多個GPU、一或多個記憶體裝置、及裝置至裝置或網路或結構介面的獨立圖形卡。並行處理單元202的不同實例可被組態成相互操作，即使不同的實例具有不同數目的處理核心、不同數量的本地並行處理器記憶體及/或其他組態差異。可選地，並行處理單

元 202 的一些實例可相對在其他實例包括較高精確度的浮點單元。結合並行處理單元 202 或並行處理器 200 中的一或多個實例的系統可用多種組態及形狀因數來實現，包括但不限於桌上型電腦、筆記型電腦或手持式個人電腦、伺服器、工作站、遊戲主控台及 / 或嵌入式系統。協調器 (orchestrator) 可以使用下列中的一或多者來形成用於工作負載效能的複合節點：分解的處理器資源，快取資源，記憶體資源、儲存資源和網路資源。

【0072】圖 2B 係分割單元 220 的方塊圖。分割單元 220 為圖 2A 的分割單元 220A-220N 中之一者的實例。如圖所示，分割單元 220 包括 L2 快取 221、框緩衝器介面 225 及 ROP 226 (柵格操作單元)。L2 快取 221 為讀取 / 寫入快取，其被組態成執行從記憶體交叉開關 216 及 ROP 226 所接收之載入及儲存操作。讀取失誤與緊急寫回請求係由 L2 快取 221 輸出至框緩衝器介面 225 以供處理。更新還可經由框緩衝器介面 225 來傳送至框緩衝器以供處理。在一實施例中，框緩衝器介面 225 係與並行處理器記憶體中的記憶體單元中之一者介接，諸如圖 2A 的記憶體單元 224A-224N (例如，在並行處理器記憶體 222 內)。分割單元 220 亦可額外地或替代地與在並行處理器記憶體中記憶體單元中的一者經由記憶體控制器 (未圖示) 介接。

【0073】在圖形應用中，ROP 226 為執行諸如模板、z 測試、混合等等柵格操作的處理單元。ROP 226 接著輸出被儲存在圖形記憶體中之經處理的圖形資料。在一些實施

例中，ROP 226包括CODEC 227或與CODEC 227耦接，包括壓縮邏輯以壓縮寫入記憶體或L2快取221的深度或顏色資料，並解壓縮從記憶體或L2快取221讀取的深度或顏色資料。所述壓縮邏輯可以是使用多種壓縮演算法中的一或多種的無損壓縮邏輯。由CODEC 227執行的壓縮類型可基於待壓縮資料的統計特性來變化。例如，在一實施例中，增量顏色壓縮係在每個磚片的基礎上，對於深度和彩色資料執行。在一實施例中，CODEC 227包括壓縮和解壓縮邏輯，其可以壓縮或解壓縮與機器學習操作相關聯的計算資料。CODEC 227可(例如)壓縮稀疏矩陣資料以進行稀疏機器學習操作。CODEC 227亦可壓縮用稀疏矩陣格式(例如，坐標列表編碼(COO)、壓縮稀疏列(CSR)、壓縮稀疏行(CSC)等)編碼的稀疏矩陣資料，以產生壓縮和編碼的稀疏矩陣資料。壓縮和編碼的稀疏矩陣資料可以在被處理元件處理之前解壓縮及/或解碼，或者處理元件可以被組態以消耗壓縮、編碼或壓縮和編碼的資料以進行處理。

【0074】ROP 226被包括在每一處理叢集(例如，圖2A的叢集214A-214N)內而非在分割單元220內。在此類實施例中，針對像素資料之讀取及寫入請求透過記憶體交叉開關216，而非像素片段資料被傳輸。經處理的圖形資料可被顯示在顯示裝置(諸如圖1的一或多個顯示裝置110之一)上、被發送以供藉由處理器102之進一步處理或被發送以供圖2A之並行處理器200內的處理實體之一的進一步處理。

【0075】圖 2C 係並行處理單元內的處理叢集 214 的方塊圖。例如，處理叢集為圖 2A 的處理叢集 214A-214N 之一的實例。處理叢集 214 可被組態成並行地執行許多執行緒，其中術語「執行緒」係指稱執行於特定組輸入資料上之特定程式的實例。可選地，單一指令、多重資料 (SIMD) 指令發送技術被用來支持大量執行緒的並行執行而不需要提供多個獨立的指令單元。替代地，單一指令、多重執行緒 (SIMT) 技術被用來支持大量一般同步化執行緒的並行執行，其使用組態成將指令發送至處理叢集之各者內的一組處理引擎的常見指令單元。不同於 SIMD 執行體系 (其中所有處理引擎通常執行相同的指令)，SIMT 執行容許不同的執行緒透過既定的執行緒程式而更輕易地跟隨分散的執行路徑。本領域技術人員將瞭解 SIMD 處理體系代表 SIMT 處理體系之功能性子集。

【0076】處理叢集 214 之操作可經由管線管理器 232 而被控制，管線管理器 232 將處理工作分配至 SIMT 並行處理器。管線管理器 232 接收來自圖 2A 的排程器 210 的指令，並經由圖形多處理器 234 及 / 或紋理單元 236 來管理那些指令之執行。所示之圖形多處理器 234 為 SIMT 並行處理器的例示性實例。然而，不同架構的各種類型 SIMT 並行處理器可被包括在處理叢集 214 內。圖形多處理器 234 中的一或多個實例可被包括在處理叢集 214 內。圖形多處理器 234 可處理資料，而資料交叉開關 240 可被用來將經處理資料分配至多個可能的目的地之一，包括其他著色器單元。管線管

理器 232 可藉由指明將經由資料交叉開關 240 而被分配之經處理資料的目的地以協助經處理資料之分配。

【0077】處理叢集 214 內的各圖形多處理器 234 可包括相同組的功能性執行邏輯(例如，算術邏輯單元、載入-儲存單元，等等)。功能性執行邏輯可用管線方式來組態，其中新的指令可在先前指令完成之前被發送。功能性執行邏輯支持多種操作，包括整數和浮點算術、比較操作、布林操作、位元移位及各種代數函數之計算。相同的功能性單元硬體可被用來執行不同的操作，且可能存在功能性單元的任何組合。

【0078】被傳輸至處理叢集 214 之指令構成執行緒。跨越該組並行處理引擎來執行的一組執行緒為執行緒群組。執行緒群組係對於不同的輸入資料執行相同的程式。執行緒群組內的各執行緒可被指派給圖形多處理器 234 內的不同處理引擎。執行緒群組可包括比圖形多處理器 234 內的處理引擎數量更少的執行緒。當執行緒群組包括比處理引擎數量更少的執行緒時，在執行緒群組正被處理之週期期間，處理引擎中的一或多者可以是閒置的。執行緒群組還可包括比圖形多處理器 234 內的處理引擎數量更多的執行緒。當執行緒群組包括比圖形多處理器 234 內的處理引擎數量更多的執行緒時，處理可在連續的時脈週期被執行。可選地，多個執行緒群組可被同時地執行於圖形多處理器 234。

【0079】圖形多處理器 234 包括用以執行載入及儲存

操作的內部快取記憶體。可選地，圖形多處理器 234 可放棄內部快取並使用處理叢集 214 內的快取記憶體(例如，第一階(L1)快取 248)。每一圖形多處理器 234 也有權存取分割單元(例如，圖 2A 的分割單元 220A-220N)內的第二階(L2)快取，該些 L2 快取可被共享於所有處理叢集 214 之間且可被用來轉移執行緒之間的資料。圖形多處理器 234 還可存取晶片外全域記憶體，其可包括本地並行處理器記憶體及/或系統記憶體中的一或多者。並行處理單元 202 外部的任何記憶體可被使用為全域記憶體。其中處理叢集 214 包括圖形多處理器 234 的多個實例的實施例可共享可被儲存於 L1 快取 248 中的共同指令及資料。

【0080】每一處理叢集 214 可包括 MMU 245(記憶體管理單元)，其被組態成將虛擬位址映射成實體位址。在其他實施例中，MMU 245 中的一或多個實例可駐存於圖 2A 的記憶體介面 218 內。MMU 245 包括被用來將虛擬位址映射成磚片之實體位址的一組頁面表項目(PTE)，及選擇性地快取線指標。MMU 245 可包括位址轉換後備緩衝(TLB)或可駐存於圖形多處理器 234 內的快取或 L1 快取或處理叢集 214。實體位址被處理以局部性地分配表面資料存取，以允許有效率的請求分區單元之間的交錯。快取線索引可被用來判定針對快取線之請求為命中或未命中。

【0081】在圖形及計算應用中，處理叢集 214 可被組態以致使各圖形多處理器 234 被耦接至紋理單元 236 以供執行紋理映射操作，例如，判定紋理樣本位置、讀取紋理資

料及過濾紋理資料。紋理資料係從內部紋理 L1 快取(未圖示)或者(在一些實施例)從圖形多處理器 234 內的 L1 快取讀取，且依照所需從 L2 快取、本地並行處理器記憶體或系統記憶體提取。每一圖形多處理器 234 將經處理的工作輸出至資料交叉開關 240 以將該經處理的工作提供至另一處理叢集 214 以供進一步處理，或者經由記憶體交叉開關 216 將該經處理的工作儲存於 L2 快取、本地並行處理器記憶體或系統記憶體中。preROP 242(預柵格操作單元)被組態成從圖形多處理器 234 接收資料、導引資料至 ROP 單元，其可連同如本文中所述之分割單元(例如，圖 2A 的分割單元 220A-220N)被配置。preROP 242 單元可執行針對顏色混合之優化、組織像素顏色資料及執行位址轉換。

【0082】應理解到本文中所示之核心架構是說明性的，並且變化和修改是可能的。任何數目的處理單元(例如，圖形多處理器 234、紋理單元 236、preROP 242，等等)可被包括在處理叢集 214 內。此外，雖然僅繪示一個處理叢集 214，但如本文中所述之並行處理單元可包括任何數量的實例的處理叢集 214。可選地，每一處理叢集 214 可被組態成使用獨立且不同的處理單元、L1 快取、L2 快取等，獨立於其他處理叢集 214 操作。

【0083】圖 2D 顯示圖形多處理器 234 的實例，其中圖形處理器 234 與處理叢集 214 之管線管理器 232 耦接。圖形多處理器 234 具有執行管線，其包括但不限於指令快取 252、指令單元 254、位址映射單元 256、暫存器檔案 258、

一或多個通用圖形處理單元(GPGPU)核心 262 及一或多個載入/儲存單元 266。GPGPU 核心 262 及載入/儲存單元 266 係經由記憶體及快取互連 268 與快取記憶體 272 及共享記憶體 270 耦接。圖形多處理器 234 可額外地包括張量及/或射線追蹤核心 263，其包括用以加速矩陣及/或射線追蹤操作的硬體邏輯。

【0084】指令快取 252 可接收用以從管線管理器 232 執行的指令串。該些指令被快取在指令快取 252 中且被調度以供指令單元 254 執行。指令單元 254 可調度指令為執行緒群組(例如，經線)，其中執行緒群組中之各執行緒被指派給 GPGPU 核心 262 內的不同執行單元。指令可藉由指明統一位址空間之中的位址來存取本地、共享或全域位址空間中的任一者。位址映射單元 256 可被用來將統一位址空間中之位址轉換為可由載入/儲存單元 266 存取的相異記憶體位址。

【0085】暫存器檔案 258 提供用於圖形多處理器 234 之功能性單元的一組暫存器。暫存器檔案 258 提供暫時儲存給連接至圖形多處理器 234 之功能性單元(例如，GPGPU 核心 262、載入/儲存單元 266)的資料路徑的操作元。暫存器檔案 258 被劃分在功能性單元的各者之間以致使各功能性單元被配置暫存器檔案 258 之專用部分。例如，暫存器檔案 258 被劃分在正由圖形多處理器 324 所執行的不同經線之間。

【0086】GPGPU 核心 262 可各包括浮點單元(FPU)及/

或整數算術邏輯單元(ALU)，其被用來執行圖形多處理器 234 之指令。在一些實施方案中，GPGPU 核心 262 可包括硬體邏輯，其可駐留張量及/或射線追蹤核心 263。GPGPU 核心 262 可於架構上類似或可於架構上不同。例如以及在一個實施例中，GPGPU 核心 262 的第一部分包括單一精確度 FPU 及整數 ALU，而 GPGPU 核心的第二部分包括雙精確度 FPU。可選地，FPU 可實現 IEEE 754-2008 標準於浮點算術或者致使可變精確度浮點算術。圖形多處理器 324 可額外包括一或多個固定功能或特殊功能單元，用以執行諸如複製矩形或像素混合操作的特定功能。GPGPU 核心中的一或多者亦可包括固定或特殊功能邏輯。

【0087】GPGPU 核心 262 包含能夠對於多組資料執行單一指令的 SIMD 邏輯。可選地，GPGPU 核心 262 可以實體地執行 SIMD4、SIMD8 和 SIMD16 指令，並且邏輯地執行 SIMD1、SIMD2 和 SIMD32 指令。用於 GPGPU 核心的 SIMD 指令可在編譯時由著色器編譯器產生或在執行程式編寫時自動產生，並且針對單一程式多重資料(SPMD)或 SIMT 架構來編譯。針對 SIMT 執行模型配置的程式的多個執行緒可以經由單一 SIMD 指令來執行。例如並且在一個實施例中，執行相同或類似操作的八個 SIMT 執行緒可以經由單一 SIMD8 邏輯單元並行地執行。

【0088】記憶體及快取互連 268 為互連網路，其將圖形多處理器 234 的功能性單元之各者連接至暫存器檔案 258 以及連接至共享記憶體 270。例如，記憶體及快取互連 268

為交叉開關互連，其容許載入/儲存單元266實現介於共享記憶體270與暫存器檔案258之間的載入和儲存操作。暫存器檔案258可操作於如GPGPU核心262之相同頻率，因此介於GPGPU核心262與暫存器檔案258之間的資料轉移是非常低延遲時間的。共享記憶體270可被用來致使介於執行在圖形多處理器234內的功能性單元上的執行緒之間的通訊。快取記憶體272可被使用為資料快取，例如，用以快取在功能性單元與紋理單元236之間傳遞的紋理資料。共享記憶體270還可被使用為程式管理快取的。共享記憶體270及快取記憶體272可與資料交叉開關240耦接以致使與處理叢集之其他組件通訊。除了被儲存於快取記憶體272內的自動快取資料以外，執行於GPGPU核心262上之執行緒可將資料藉由程式化儲存於共享記憶體內。

【0089】圖3A-3C繪示根據實施例之額外的圖形多處理器。圖3A-3B繪示圖形多處理器325、350，其與圖2C之圖形多處理器234有關，並且可替代其中之一者使用。因此，與圖形多處理器234結合之任何特徵的揭露在本文中亦可揭露與圖形多處理器325、350的相應結合，但不限於此。圖3C繪示圖形處理單元(GPU)380，其包括配置成對應於圖形多處理器325、350的多核心群組365A-365N的專用圖形處理資源組。所示之圖形多核心處理器325、350及多核心群組365A-365N可以是能夠同時執行大量執行緒之串流多處理器(SM)。

【0090】圖3A的圖形多處理器325相對於圖2D的圖形

多處理器 234 包括執行資源單元的多個額外實例。例如，圖形多處理器 325 可包括指令單元 332A-332B、暫存器檔案 334A-334B 及紋理單元 344A-344B 的多個實例。圖形多處理器 325 還包括多組圖形或計算執行單元(例如，GPGPU 核心 336A-336B、GPGPU 核心 337A-337B、GPGPU 核心 338A-338B)及多組載入/儲存單元 340A-340B。執行資源單元具有共同指令快取 330、紋理及/或資料快取記憶體 342 及共享記憶體 346。

【0091】各個組件可經由互連結構 327 來通訊。互連結構 327 包括一或多個交叉開關，用以致使介於圖形多處理器 325 的各個組件之間的通訊。互連結構 327 是圖形多處理器 325 的每個組件堆疊在其上的獨立、高速網路構造層。圖形多處理器 325 的組件經由互連結構 327 與遠端組件通訊。例如，核心 336A-336B、337A-337B 和 3378A-338B 可各自經由互連結構 327 與共享記憶體 346 進行通訊。互連結構 327 可仲裁圖形多處理器 325 內的通訊，以確保組件之間的公平頻寬分配。

【0092】圖 3B 圖形多處理器 350 包括多組執行資源 356A-356D，其中各組執行資源包括多個指令單元、暫存器檔案、GPGPU 核心及載入儲存單元，如圖 2D 及圖 3A 中所示。執行資源 356A-356D 可配合紋理單元 360A-360D 工作以供紋理操作，同時共享指令快取 354 及共享記憶體 353。例如，執行資源 356A-356D 可共享指令快取 354 和共享記憶體 353，以及紋理及/或資料快取記憶體 358A-358B

的多個實例。各個組件可經由類似於圖 3A 之互連結構 327 的互連結構 352 來通訊。

【0093】本領域技術人員將理解，圖 1、2A-2D 及 3A-3B 中所述之架構為說明性的而不是限制本實施例的範圍。因此，本文中所述之技術可被實現於任何適當組態的處理單元上，包括(非限制)一或多個行動應用程式處理器、包括多核心 CPU 的一或多個桌上型電腦或伺服器中央處理單元(CPU)、一或多個並行處理單元(諸如圖 2A 的並行處理單元 202)，以及一或多個圖形處理器或專用處理單元，而不背離本文中所述之實施例的範圍。

【0094】如本文中所述之並行處理器或 GPGPU 被通訊地耦接至主機/處理器核心以加速圖形操作、機器學習操作、模式分析操作及各種通用 GPU(GPGPU)功能。GPU 可透過匯流排或其他互連(例如，諸如 PCIe 或 NVLink 等高速互連)被通訊地耦接至主機處理器/核心。於其他實施例中，GPU 可被集成於如該等核心之相同封裝或晶片上，且透過內部處理器匯流排/互連(亦即，封裝或晶片內部)而被通訊地耦接至該些核心。無論 GPU 被連接的方式為何，處理器核心可依工作描述符所含之命令/指令的序列之形式將工作分配給 GPU。GPU 接著使用專用電路/邏輯以便有效率地處理這些命令/指令。

【0095】圖 3C 繪示圖形處理單元(GPU)380，其包括配置為多核心群組 365A-365N 的專用圖形處理資源組。雖然僅提供了單個多核心群組 365A 的細節，但是應當理解到

其他多核心群組 365B-365N 可以配備有相同或相似的圖形處理資源組。關於多核心群組 365A-365N 敘述的細節也可應用於本文所述之任何圖形多處理器 234、325、350。

【0096】如所示，多核心群組 365A 可包括圖形核心 370 組、張量核心 371 組、及射線追蹤核心 372 組。排程器/調度器 368 排程並調度圖形執行緒，以在各種核心 370、371、372 上執行。暫存器檔案 369 組當執行圖形執行緒時儲存由核心 370、371、372 使用的操作元值。這些可包括(例如)用於儲存整數值的整數暫存器、用於儲存浮點的浮點暫存器、用於儲存封裝資料元件(整數及/或圖點資料元件)的向量暫存器、及儲存張量/矩陣值的磚片暫存器。磚片暫存器可實施為組合的向量暫存器組。

【0097】一或多個組合第一階(L1)快取和共享記憶體單元 373 在每一多核心群組 365A 內本地儲存圖形資料，諸如紋理資料、頂點資料、像素資料、射線資料、邊界體積資料等等。可使用一或多個紋理單元 374 以執行紋理操作，諸如紋理映射及取樣。由多核心群組 365A-365N 的全部或子集共享的第 2 階(L2)快取 375 儲存圖形資料及/或多個並行圖形執行緒的指令。如所繪示，可以在複數個多核心群組 365A-365N 之間共享 L2 快取 375。一或多個記憶體控制器 367 將 GPU 380 耦接至記憶體 366，其可以是系統記憶體(例如，DRAM)及/或專用圖形記憶體(例如，GDDR6 記憶體)。

【0098】輸入/輸出(I/O)電路 363 將 GPU 380 耦接到一

或多個 I/O 裝置 362，諸如數位訊號處理器 (DSP)、網路控制器、或使用者輸入裝置。晶載互連可用於將 I/O 裝置 362 耦接至 GPU 380 和記憶體 366。I/O 電路 363 的一或多個 I/O 記憶體管理單元 (IOMMU) 364 將 I/O 裝置 362 直接耦接至系統記憶體 366。可選地，IOMMU 364 管理多組頁面表，以將虛擬位址映射至系統記憶體 366 中的實體位址。然後，I/O 裝置 362、CPU 361、和 GPU 380 可以共享相同的虛擬位址空間。

【0099】在 IOMMU 364 的實施方式中，IOMMU 364 支持虛擬化。在此情況下，它可以管理用於將訪客/圖形虛擬位址映射到訪客/圖形實體位址的第一組頁面表，以及用於將訪客/圖形實體位址映射到系統/主機實體位址的第二組頁面表(例如，在系統記憶體 366 中)。第一和第二組頁面表中每一者的基底位址可以將儲存在控制暫存器中並在上下文開關上換出(例如，以便為新上下文提供對相關頁面表組的存取)。儘管在圖 3C 中未繪示，但是核心 370、371、372 及/或多核群組 365A-365N 中的每一個都可以包括轉譯後備緩衝器 (TLB)，以快取訪客虛擬至訪客實體轉譯、訪客實體至主機實體轉譯、及訪客虛擬至主機實體轉譯。

【0100】CPU 361、CPU 380 及 I/O 裝置 362 可整合於單一半導體晶片及/或晶片封裝上。所繪示之記憶體 366 可被整合在相同晶片上或可經由晶片外介面耦接至記憶體控制器 367。在一個實施方式中，記憶體 366 包含 GDDR6 記憶

體，其可與其他實體系統級記憶體共享相同的虛擬位址空間，儘管本文敘述的基本原理不限於該特定實施方式。

【0101】張量核心 371 可包括專門設計用於執行矩陣操作的複數個執行單元，矩陣操作是用於執行深度學習操作的基本計算操作。例如，同時矩陣乘法操作可用於類神經網路訓練和推理。張量核心 371 可以使用各種操作元精度來執行矩陣處理，包括單精度浮點數(例如，32 位元)、半精度浮點數(例如，16 位元)、整數字(16 位元)、位元組(8 位元)及半位元組(4 位元)。例如，類神經網路實施方式提取每一渲染場景的特徵，並潛在地從多個訊框中組合細節，以構建高品質的最終影像。

【0102】在深度學習實施方式中，可排程並行矩陣乘法工作以在張量核心 371 上執行。類神經網路的訓練特別需要大量的矩陣點乘積操作。為了處理乘以 $N \times N \times N$ 矩陣的內積公式，張量核心 371 可以包括至少 N 個點乘積處理元件。在矩陣乘法開始之前，將一個完整矩陣加載到磚片暫存器中，並在第二矩陣的至少一個行係在每一週期加載，加載 N 個週期。每個週期有 N 個點乘積被處理。

【0103】取決於特定的實現方式，可以以不同的精度儲存矩陣元素，包括 16 位元字、8 位元位元組(例如，INT8)、4 位元半位元組(例如，INT4)。可以為張量核心 371 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，諸如可以將量化容忍到位元組和半位元組的推理工作負載)。支持的格式額外地包括 64 位元

浮點 (FP64) 和非 IEEE 浮點格式，諸如 bfloat16 格式 (例如，Brain 浮點)、具有一個符號位元、八個指數位元及八個有效位元的 16 位元浮點格式，其中明確儲存了 7 個。一個實施例包括對精度降低的張量浮點格式 (TF32) 的支持，該格式具有 FP32 (8 位元) 與 FP16 (10 位元) 的精度範圍。可以在 FP32 輸入上執行精度降低的 TF32 操作，並以相對於 FP32 更高的效能和相對於 FP16 更高的精度產生 FP32 輸出。

【0104】在一實施例中，張量核心 371 支持矩陣的稀疏操作模式，其中絕大多數值是零。張量核心 371 包括對於以稀疏矩陣表示 (例如，坐標列表編碼 (COO)、壓縮稀疏列 (CSR)、壓縮稀疏行 (CSC) 等) 編碼的稀疏輸入矩陣的支持。在稀疏矩陣表示可以被進一步壓縮的情況下，張量核心 371 還包括對壓縮稀疏矩陣表示的支持。張量核心 371 可以準備壓縮、編碼及 / 或壓縮和編碼的矩陣資料以及相關聯的壓縮及 / 或編碼元資料，並且可以提取非零值。例如，對於給定的輸入矩陣 A，可以從矩陣 A 的至少一部分的壓縮及 / 或編碼表示中加載非零值。基於可以從與非零值相關聯的索引或座標元資料判定之非零值的矩陣 A 中的位置，可以加載輸入矩陣 B 中的對應值。取決於要執行的操作 (例如，乘法)，如果對應值是零值，則可以旁路來自輸入矩陣 B 之值的加載。在一實施例中，可以藉由排程器邏輯來預掃描諸如乘法操作之類的某些操作之值對，並且僅排程非零輸入之間的操作。取決於矩陣 A 和矩陣 B 的尺寸以及要執行的操作，輸出矩陣 C 可以是密集的或稀疏

的。在輸出矩陣 C 是稀疏的情況下，並且取決於張量核心 371 的組態，可以以壓縮格式、稀疏編碼、或壓縮稀疏編碼來輸出輸出矩陣 C。

【0105】射線追蹤核心 372 可以針對實時射線追蹤和非實時射線追蹤實現方式加速射線追蹤操作。特別地，射線追蹤核心 372 可包括射線遍歷/相交電路，用於使用邊界體積等級 (BVH) 來執行射線遍歷並識別在射線與包圍在 BVH 體積內的基元之間的相交。射線追蹤核心 372 亦可包括用於執行深度測試和剔除 (例如，使用 Z 緩衝器或類似配置) 的電路。在一實施方式中，射線追蹤核心 372 與本文所述的影像去噪技術一致地執行遍歷和相交操作，其中至少一部分可以在張量核心 371 上執行。例如，張量核心 371 可以實施深度學習類神經網路以對由射線追蹤核心 372 產生的訊框執行去噪。然而，CPU 361、圖形核心 370 及/或射線追蹤核心 372 也可以實現全部或部分去噪及/或深度學習算法。

【0106】另外，如上所述，可以採用分散式去噪方法，其中 GPU 380 位於透過網路或高速互連耦接至其他計算裝置的計算裝置中。在這種分散式方法中，互連的計算裝置可以共享類神經網路學習/訓練資料，以提高整個系統學習針對不同類型的影像框及/或不同圖形應用執行去噪的速度。

【0107】射線追蹤核心 372 可處理所有的 BVH 遍歷及/或射線基元交點，從而避免圖形核心 370 因每條射線成千

上萬的指令而過載。例如，每一射線追蹤核心 372 包括第一組專用電路，用於執行邊界框測試(例如，用於遍歷操作)及/或第二組專用電路，用於執行射線-三角形相交測試(例如，相交的射線，已遍歷)。因此，例如，多核心群組 365A 可簡單地發射射線探測器，並且射線追蹤核心 372 獨立地執行射線遍歷和相交並且將命中資料(例如，命中、無命中、多次命中等)返回給執行緒上下文。當射線追蹤核心 372 執行遍歷和相交操作時，其他核心 370、371 被釋放以執行其他圖形或計算工作。

【0108】可選地，每一射線追蹤核心 372 可以包括執行 BVH 測試操作的遍歷單元及/或執行射線基元相交測試的相交單元。相交單元產生「命中」、「無命中」、或「多次命中」回應，它提供適當的執行緒。在遍歷和相交操作期間，釋放其他核心(例如，圖形核心 370 和張量核心 371)的執行資源以執行其他形式的圖形工作。

【0109】在下文敘述的一個可選實施例中，使用混合光柵化/射線追蹤方法，其中工作分佈在圖形核心 370 和射線追蹤核心 372 之間。

【0110】射線追蹤核心 372(及/或其他核心 370、371)可以包括對射線追蹤指令集的硬體支持，諸如 Microsoft 的 DirectX 射線追蹤(DXR)，其包括 DispatchRays 命令，以及射線產生、最接近命中、任何命中及錯過的著色器，可為每一物件分配唯一的一組著色器和紋理。可以由射線追蹤核心 372、圖形核心 370 及張量核心 371 支持的另一個射線

追蹤平台是 Vulkan 1.1.85。但是，請注意，本文所述的基本原理不限於任何特定的射線追蹤ISA。

【0111】一般而言，各種核心 372、371、370 可支持包括用於射線產生、最接近命中、任何命中、射線基元相交、每一基元和分層邊界框構造、錯過、訪問和例外中之一或多個指令和功能的射線追蹤指令集。更具體地，優選實施例包括執行以下一或多個功能的射線追蹤指令：

【0112】射線產生-可以為每一像素、樣本或其他使用者界定的工作分配執行射線產生指令。

【0113】最接近命中-可以執行最接近命中指令來定位射線與場景中基元的最接近交點。

【0114】任何命中-任何命中指令可標識射線與場景中的基元之間的多個交點，從而有可能標識新的最接近交點。

【0115】相交-相交指令執行射線基元相交測試並輸出結果。

【0116】每一基元邊界框構造-該指令圍繞給定的基本邊界或一組基本邊界構建邊界框(例如，在構建新的BVH或其他加速資料結構時)。

【0117】錯過-表示射線錯過場景或場景指定區域內的所有幾何圖形。

【0118】訪問-表示射線將遍歷的子體積。

【0119】異常-包括各種類型的異常處理處置(例如，針對各種錯誤條件調用)。

【0120】在一實施例中，射線追蹤核心372可以適於加速通用計算操作，其可使用類似於射線相交測試的計算技術來加速。可以提供一個計算框架，該框架使著色器程式可以經由射線追蹤核心被編譯為執行通用計算操作的低階指令及/或基元。可以從在射線追蹤核心372上執行的計算操作中受益的例示性計算問題包括涉及在坐標空間內的波束、波、射線或粒子傳播的計算。可以相對於坐標空間內的幾何形狀或網格計算與該傳播關聯的相互作用。例如，可以透過使用經由射線追蹤核心執行的指令或基元來加速與通過環境的電磁訊號傳播相關聯的計算。可以將物件在環境中對訊號的繞射和反射計算為直接射線追蹤類比。

【0121】射線追蹤核心372也可以用於執行與射線追蹤不直接相似的計算。例如，可以使用射線追蹤核心372來加速網格投影、網格細化和體積採樣計算。也可以執行通用坐標空間計算，例如最近鄰計算。例如，可以藉由在點周圍的坐標空間中界定邊界框來發現給定點附近的點集。然後，可以使用射線追蹤核心372中的BVH和射線探測器邏輯來判定邊界框中的點集。相交構成了原點，並且該原點最近。可以與在圖形核心370和張量核心371上執行的計算並行地執行使用射線追蹤核心372執行的計算。著色器編譯器可組態以將計算著色器或其他通用圖形處理程式編譯為可以在圖形核心370、張量核心371和射線追蹤核心372之間並行化的低階基元。

用於 GPU 至主處理器互連之技術

【0122】圖 4A 繪示其中複數個 GPU 410-413 (諸如圖 2A 中所示之並行處理器 200) 係透過高速鏈路 440A-440D (例如，匯流排、點對點互連，等等) 通訊地耦接至複數個多核心處理器 405-406 例示性架構。高速鏈路 440A-440D 根據實現支持 4GB/s、30GB/s、80GB/s 或更高的通訊通量。各種互連協定可被使用，包括但不限於 PCIe 4.0 或 5.0 及 NVLink 2.0。然而，本文敘述之基本原理不限於任何特定的通訊協定或通量。

【0123】GPU 410-413 之二或更多者可透過高速鏈路 442A-442B 而被互連，其可使用相同於或不同於那些用於高速鏈路 440A-440D 的協定/鏈路來實施。類似地，多核心處理器 405-406 中的二或更多者可透過高速鏈路 443 而被連接，高速鏈路 443 可以是操作在 20GB/s、30GB/s、120GB/s 或更高速的對稱多處理器 (SMP) 匯流排。替代地，介於圖 4A 中所示的各種系統組件之間的所有通訊可使用相同的協定/鏈路來完成 (例如，透過共同互連結構)。然而，如上所述，本文敘述之基本原理不限於任何特定類型的互連技術。

【0124】每一多核心處理器 405-406 個別地經由記憶體互連 430A-430B 被通訊地耦接至處理器記憶體 401-402，且每一 GPU 410-413 係透過 GPU 記憶體互連 450A-450D 而被個別通訊地耦接至 GPU 記憶體 420-423。記憶體互連 430A-

430B及450A-450D可利用相同或不同的記憶體存取技術。舉例而言(且非限制性)，處理器記憶體401-402及GPU記憶體420-423可以是揮發性記憶體，諸如動態隨機存取記憶體(DRAM)(包括堆疊DRAM)、圖形DDR SDRAM(GDDR)(例如，GDDR5、GDDR6)或高頻寬記憶體(HBM)及/或可以是諸如3D XPoint/Optane或Nano-Ram的非揮發性記憶體。例如，記憶體之某部分可以是揮發性記憶體而其他部分可以是非揮發性記憶體(例如，使用第二階記憶體(2LM)階層)。本文所述至記憶體子系統可以與多種記憶體技術兼容，諸如由JEDEC(聯合電子設備工程委員會)發布的雙倍資料速率版本。

【0125】如以下所述，雖然各個處理器405-406及GPU 410-413可被個別地實體地耦接至特定記憶體401-402、420-423，可實現一種統一記憶體架構，其中相同的虛擬系統位址空間(也稱為「有效位址」空間)被分配於所有的各個實體記憶體之間。例如，處理器記憶體401-402可各包含64GB的系統記憶體位址空間，而GPU記憶體420-423可各包含32GB的系統記憶體位址空間(在此實例中，導致總共256GB的可定址記憶體)。

【0126】圖4B繪示針對介於多核心處理器407與圖形加速模組446之間的互連的額外光學細節。圖形加速模組446可包括整合於經由高速鏈路440被耦接至處理器407之線卡上的一或多個GPU晶片。替代地，圖形加速模組446可被整合於如處理器407的相同封裝或晶片上。

【0127】所繪示的處理器407包括複數個核心460A-460D，其各具有轉換後備緩衝器461A-461D及一或多個快取462A-462D。該些核心可包括用以執行指令及處理資料之各種其他組件，其未被繪示以免混淆本文所敘述的基本原理(例如，指令提取單元、分支預測單元、解碼器、執行單元、記錄器緩衝器，等等)。快取462A-462D可包含第1階(L1)及第2階(L2)快取。此外，一或多個共享快取456可被包括在快取階層中且由多組核心460A-460D所共享。例如，處理器407的一個實施例包括24個核心，各具有其本身的L1快取、十二個共享的L2快取，以及十二個共享的L3快取。在這實施例中，L2及L3快取之一係由兩個相鄰核心所共享。處理器407及圖形加速器整合模組446係與可包括處理器記憶體401-402的系統記憶體441連接。

【0128】儲存在各個快取462A-462D、456及系統記憶體441中的資料和指令係經由透過同調匯流排464之核心間通訊來維持同調性。例如，各快取可具有與其相關的快取同調性邏輯/電路，以回應於檢測到的讀取或寫入至特定快取線而透過同調匯流排464來通訊。在一種實施方式中，快取監聽協定被實現於同調匯流排464以監聽快取存取。快取監聽/同調性技術是本領域技術人員所熟知的且在此將不被詳細地描述，以免混淆本發明之基本原理。

【0129】代理電路425將圖形加速模組446通訊地耦接至同調匯流排464，其容許圖形加速模組446參與快取同調性協定而作為該些核心的同級。特別地，介面435透過高

速鏈路 440(例如，PCIe 匯流排、NVLink，等等)提供連接至代理電路 425，而介面 437 將圖形加速模組 446 連接至高速鏈路 440。

【0130】在一種實施方式中，加速器積體電路 436 為了圖形加速模組 446 的複數個圖形處理引擎 431、432、N，提供了快取管理、記憶體存取、背景管理及中斷管理服務。圖形處理引擎 431、432、N 可各包含各自的圖形處理單元 (GPU)。替代地，圖形處理引擎 431、432、N 可包含 GPU 內的不同類型的圖形處理引擎，諸如圖形執行單元、媒體處理引擎(例如，視訊編碼器/解碼器)、取樣器及分裂引擎。換言之，圖形加速模組可以是具有複數個圖形處理引擎 431-432、N 之 GPU，或者圖形處理引擎 431-432、N 可以是整合於共同封裝、線卡或晶片上之單獨 GPU。

【0131】加速器積體電路 436 包括記憶體管理單元 (MMU) 439，其用以執行各種記憶體管理功能，諸如虛擬至實體記憶體轉換(亦稱為有效至真實記憶體轉換)及用以存取系統記憶體 441 之記憶體存取協定。MMU 439 還可包括轉換後備緩衝器 (TLB)(未圖示)，其用以將虛擬/有效快取至實體/真實位址轉換。在一種實施方式中，快取 438 儲存命令及資料以供圖形處理引擎 431、432、N 有效率地存取。可以將儲存在快取 438 和圖形記憶體 433-434、M 中的資料保持與核心快取 462A-462D、456 系統記憶體 411 同調。如上所述，這可經由代理電路 425 來完成，該代理電路 425 為了快取 438 及記憶體 433-434、M 參與快取同調性機

制(例如，傳送更新至相關於處理器快取 462A-462D、456 上之快取線的修改/存取的快取 438，並從快取 438 接收更新)。

【0132】一組暫存器 445 儲存了背景資料以用於由圖形處理引擎 431-432、N 所執行之執行緒，而背景管理電路 448 係管理執行緒背景。例如，背景管理電路 448 可執行保存及復原操作，以保存及復原背景切換期間之各個執行緒的背景(例如，其中第一執行緒被保存而第二執行緒被儲存以致該第二執行緒可由圖形處理引擎來執行)。例如，在背景切換上，背景管理電路 448 可將目前的暫存器值儲存至記憶體中的指定區(例如，由背景指標所識別)。當返回至該背景時，其可接著復原該些暫存器值。中斷管理電路 447 接收並處理從系統裝置所接收的中斷。

【0133】在一種實施方式中，來自圖形處理引擎 431 之虛擬/有效位址係藉由 MMU 439 而被轉換為系統記憶體 441 中的真實/實體位址。可選地，加速器積體電路 436 的一個實施例支持多個(例如，4、8、16)圖形加速模組 446 及/或其他加速器裝置。圖形加速模組 446 可專用於處理器 407 上所執行的單一應用程式或者可被共享於多個應用程式之間。可選地，提供了虛擬化的圖形執行環境，其中圖形處理引擎 431-432、N 之資源係與多個應用程式、虛擬機器(VM)、或容器共享。該些資源可被細分為「片」，其根據與 VM 及/或應用程式相關的處理需求和優先權而被配置至不同的 VM 及/或應用程式。VM 和容器在本文中可

互換使用。

【0134】虛擬機(VM)可以是運行作業系統及一或多個應用程式的軟體。VM可由規範、組態檔案、虛擬磁碟檔、非揮發性隨機存取記憶體(NVRAM)設定檔、及日誌檔界定，並且由主機計算平台的實體資源支持。VM可以包括安裝在模擬專用硬體的軟體上的作業系統(OS)或應用程式環境。終端使用者在虛擬機上的體驗與專用硬體上的體驗相同。稱為超管理器的專用軟體可以完全模擬PC客戶端或伺服器的CPU、記憶體，硬碟、網路及其他硬體資源，從而使虛擬機可以共享資源。超管理器可模擬彼此隔離的多個虛擬硬體平台，從而允許虛擬機在同一基礎實體主機上運行Linux®、Windows®Server、VMware ESXi和其他作業系統。

【0135】容器可以是應用程式、組態和相依性的軟體封裝，因此應用程式可以在一個計算環境中可靠地運行到另一個計算環境。容器可以共享安裝在伺服器平台上的作業系統，並如隔離的程序運行。容器可為軟體封裝，其含有軟體需要運行的所有內容，諸如系統工具、庫及設定。容器的安裝方式不像傳統的軟體程式那樣，其可以使它們與其他軟體和作業系統本身隔離。容器的隔離性質提供了許多好處。首先，容器中的軟體將在不同的環境中運行相同的軟體。例如，包括PHP及MySQL的容器可以在Linux®電腦和Windows®電腦上相同地運行。其次，容器提供了額外的安全性，因為該軟體不會影響主機作業系統。儘管

已安裝的應用程式可能會更改系統設置並修改資源(諸如，Windows註冊表)，但是容器只能修改容器內的設置。

【0136】因此，加速器積體電路436係作用為針對圖形加速模組446的對於系統的橋接，並提供位址轉換及系統記憶體快取服務。在一實施例中，為了促進橋接功能，加速器積體電路436亦可包括共享的I/O 497(例如，PCIe、USB或其他)和硬體，以實現對電壓、時序、效能、散熱和安全性的系統控制。共享的I/O 497可以利用單獨的實體連接，或者可以遍歷高速鏈路440。此外，加速器積體電路436可提供虛擬化協助給主機處理器，以管理圖形處理引擎之虛擬化、中斷及記憶體管理。

【0137】因為圖形處理引擎431-432、N之硬體資源被明確地映射至由主機處理器407所見之真實位址空間，所以任何主機處理器可使用有效位址值來直接地定址這些資源。加速器積體電路436的一個可選的功能為圖形處理引擎431-432、N之實體分離，以致使其對系統而言呈現為獨立的單元。

【0138】一或多個圖形記憶體433-434、M被個別地耦接至圖形處理引擎431-432、N之各者。圖形記憶體433-434、M儲存由圖形處理引擎431-432，N之各者所處理的指令及資料。圖形記憶體433-434、M可以是揮發性記憶體，諸如DRAM(包括堆疊DRAM)、GDDR記憶體(例如，GDDR5、GDDR6)或HBM，及/或可以是非揮發性記憶體，諸如3D XPoint/Optane、Samsung Z-NAND或Nano-Ram。

【0139】為了減少透過高速鏈路440之資料流量，偏移技術被用來確保圖形記憶體433-434、M中所儲存之資料為將最常被圖形處理引擎431-432、N使用且較佳地不被核心460A-460D使用(至少不常)之資料。類似地，偏移機制嘗試將該些核心(且較佳地不是圖形處理引擎431-432、N)所需的資料保持於該些核心及系統記憶體411的快取462A-462D、456內。

【0140】根據圖4C繪示的變體，加速器積體電路436被整合於處理器407內。圖形處理引擎431-432、N透過高速鏈路440經由介面437及介面435(其再次地，可利用任何形式的匯流排或介面協定)直接地通訊至加速器積體電路436。加速器積體電路436可執行如關於圖4B所述那些的相同操作，但由於其接近於同調匯流排464及快取462A-462D、456潛在地具有較高的產量。

【0141】所述實施例支持包括(無圖形加速模組虛擬化的)專用程序編程模型及(具有虛擬化的)共享編程模型的不同編程模型。後者可包括由加速器積體電路436所控制的編程模型以及由圖形加速模組446所控制的編程模型。

【0142】在專用程序模型的一個實施例中，圖形處理引擎431、432、... N可以係專用於單一作業系統下的單一應用程式或程序。單一應用程式可將其他應用程式請求注入至提供VM/分割內的虛擬化的圖形引擎431、432、... N。

【0143】在專用程序編程模型中，圖形處理引擎

431、432、N可由多個VM/應用程式分割所共享。該些共享模型需要系統超管理器來將圖形處理引擎431-432、N虛擬化以容許各作業系統的存取。針對沒有超管理器的單一分割系統，圖形處理引擎431-432、N係由作業系統所擁有。在兩種情況下，作業系統可將圖形處理引擎431-432、N虛擬化以提供對於各程序或應用程式之存取。

【0144】對於共享的編程模型，圖形加速模組446或個別圖形處理引擎431-432、N係使用程序處置來選擇程序元件。程序元件被儲存在系統記憶體441中且使用本文中所述之有效位址至真實位址轉換技術而為可定址的。程序處置可以是當以圖形處理引擎431-432、N登錄其背景時(亦即，用以將程序元件添加至程序元件鏈路列表之呼叫系統軟體)，被提供至主機程序之實現特定的值。程序處置的較低16位元可以是程序元件鏈路列表內的程序元件的偏移。

【0145】圖4D繪示例示性加速器整合片490。如本文中所使用，「片」包含加速器積體電路436的處理資源之指定部分。系統記憶體441內的應用程式有效位址空間482係儲存程序元件483。程序元件483回應於來自處理器407上所執行之應用程式480的GPU調用481而被儲存。程序元件483含有相應應用程式480之程序狀態。程序元件483中所含之工作描述符(WD)484可以是應用程式所請求的單一工作或可含有指向工作之佇列的指標。在後者的情況下，WD 484為指向應用程式之位址空間482中的工作請求佇列

之指標。

【0146】圖形加速模組 446 及 / 或個別圖形處理引擎 431-432、N 可由系統中程序的全部或子集所共享。例如，本文敘述的技術包括一種用以設定程序狀態並傳送 WD 484 至圖形加速模組 446 以在虛擬化環境中開始工作的基礎架構。

【0147】在一種實施方式中，專用程序編程模型為實現特定的。在此模型中，單一程序擁有圖形加速模組 446 或單獨圖形處理引擎 431。因為圖形加速模組 446 係由單一程序所擁有，所以在當圖形加速模組 446 被指派時的時刻，超管理器係針對其擁有的分割來將加速器積體電路 436 初始化，而作業系統係針對其擁有的程序來將加速器積體電路 436 初始化。

【0148】在操作中，加速器整合片 490 中之 WD 提取單元 491 提取了下一個 WD 484，其包括將由圖形加速模組 446 的圖形處理引擎之一者所完成的工作之指示。來自 WD 484 之資料可被儲存於暫存器 445 中並由 MMU 439、中斷管理電路 447 及 / 或背景管理電路 448 所使用。例如，MMU 439 包括用以存取 OS 虛擬位址空間 485 內的分段 / 頁面表 486 的分段 / 頁面步行電路。中斷管理電路 447 可處理從圖形加速模組 446 所接收的中斷事件 492。當執行圖形操作時，由圖形處理引擎 431-432、N 所產生的有效位址 493 被 MMU 439 轉換為真實位址。

【0149】相同組的暫存器 445 被複製給各圖形處理引

擎 431-432、N 及 / 或 圖 形 加 速 模 組 446， 並 且 可 由 超 管 理 器 或 作 業 系 統 來 初 始 化。 這 些 複 製 的 暫 存 器 之 各 者 可 被 包 括 在 加 速 器 整 合 片 490 中。 在 一 實 施 例 中， 每 一 圖 形 處 理 引 擎 431-432、N 可 以 將 其 作 為 不 同 的 圖 形 處 理 器 裝 置 呈 現 給 超 管 理 器 496。 可 以 為 特 定 圖 形 處 理 引 擎 431-432、N 的 客 戶 端 組 態 QoS 設 置， 並 且 可 以 啟 用 每 一 引 擎 的 客 戶 端 之 間 的 資 料 隔 離。 可 由 超 管 理 器 所 初 始 化 之 例 示 性 暫 存 器 係 顯 示 於 表 1 中。

表 1-超管理器初始化暫存器

1	片控制暫存器
2	真實位址(RA)排程的處理區域指標
3	權限遮罩改寫暫存器
4	中斷向量表項目偏移
5	中斷向量表項目限制
6	狀態暫存器
7	邏輯分割ID
8	真實位址(RA)超管理器加速器利用記錄指標
9	儲存描述暫存器

【 0150】 可 由 作 業 系 統 所 初 始 化 之 例 示 性 暫 存 器 係 顯 示 於 表 2 中。

表 2-作業系統初始化暫存器

1	程序及執行緒識別
2	有效位址(EA)背景保存/復原指標

3	虛擬位址(VA)加速器利用記錄指標
4	虛擬位址(VA)儲存分段表指標
5	權限遮罩
6	工作描述符

【0151】 每一 WD 484係專用於特定的圖形加速模組 446及/或圖形處理引擎 431-432、N。其含有圖形處理引擎 431-432、N欲執行其工作所需的所有資訊或者其可以是指向應用程式已設定將完成之工作的命令佇列之記憶體位置的指標。

【0152】 圖 4E繪示共享模型的額外細節。其包括程序元件列表 499儲存於其中的超管理器真實位址空間 498。超管理器真實位址空間 498為經由將作業系統 495之圖形加速模組引擎虛擬化之超管理器 496可存取的。

【0153】 共享編程模型容許來自該系統中的分割的全部或子集之程序的全部或子集使用圖形加速模組 446。有圖形加速模組 446係由多個程序及分割所共享的兩種編程模型：時間片共享及圖形導向共享。

【0154】 在此模型中，系統超管理器 496擁有圖形加速模組 446並使其功能可用於所有作業系統 495。為了使圖形加速模組 446支持藉由系統超管理器 496之虛擬化，圖形加速模組 446可遵守以下需求：1)應用程式之工作請求需為自主性的(亦即，狀態無須被維持於工作之間)；或者圖形加速模組 446需提供背景保存及復原機制。2)應用程式之工作請求係由圖形加速模組 446確保在指定的時間量中

完成，包括任何轉換錯誤，或者圖形加速模組 446 提供預估該工作之處理的能力。3) 當操作於該導向的共享編程模型中時，圖形加速模組 446 需被確保程序之間的公平性。

【0155】對於共享模型，應用程式 480 被要求利用圖形加速模組 446 類型、工作描述符(WD)、權限遮罩暫存器(AMR)值及背景保存/復原區域指標(CSRP)進行作業系統 495 系統呼叫。圖形加速模組 446 類型敘述了系統呼叫的目標加速功能。圖形加速模組 446 類型可以是系統特定值。WD 被明確地針對圖形加速模組 446 格式化並可以是以下之形式：圖形加速模組 446 命令、指向使用者界定結構之有效位址指標、指向命令佇列之有效位址指標或用以描述將由圖形加速模組 446 所完成之工作的任何其他資料結構。於一實施例中，AMR 值為用於目前程序之 AMR 狀態。被傳遞至作業系統之值係類似於設定 AMR 之應用程式。假如加速器積體電路 436 及圖形加速模組 446 實現不支持使用者權限遮罩改寫暫存器(UAMOR)，則作業系統可在超管理器呼叫中傳遞 AMR 之前，將目前的 UAMOR 值應用於 AMR 值。在將該 AMR 置於程序元件 483 中之前，超管理器 496 可選擇性地應用當前權限遮罩改寫暫存器(AMOR)值。CSRП 可為含有圖形加速模組 446 之應用程式位址空間 482 中的區域之有效位址的暫存器 445 之一者，其用以保存並復原背景狀態。假如沒有狀態需被保存於工作之間或者當工作被預估時，則此指標為選擇性的。背景保存/復原區域可以是固定的系統記憶體。

【 0156 】 在接收該系統呼叫時，作業系統 495 可驗證應用程式 480 已登錄且已被提供權限來使用圖形加速模組 446。作業系統 495 接著呼叫具有表 3 中所示之資訊的超管理器 496。

表 3-OS 至超管理器呼叫參數

1	工作描述符(WD)
2	權限遮罩暫存器(AMR)值(潛在地遮蔽的)
3	有效位址(EA)背景保存/復原區域指標(CSRP)
4	程序ID(PID)及選擇性執行緒ID(TID)
5	虛擬位址(VA)加速器利用記錄指標(AURP)
6	儲存分段表指標(SSTP)之虛擬位址
7	邏輯中斷服務數(LISN)

【 0157 】 在接收超管理器呼叫時，超管理器 496 驗證了作業系統 495 已登錄且已被提供權限來使用圖形加速模組 446。超管理器 496 接著將程序元件 483 置入用於相應圖形加速模組 446 類型的程序元件鏈路表中。程序元件可包括表 4 中所示之資訊。

表 4-程序元件資訊

1	工作描述符(WD)
2	權限遮罩暫存器(AMR)值(潛在地遮蔽的)
3	有效位址(EA)背景保存/復原區域指標(CSRP)
4	程序ID(PID)及選擇性執行緒ID(TID)
5	虛擬位址(VA)加速器利用記錄指標(AURP)
6	儲存分段表指標(SSTP)之虛擬位址
7	邏輯中斷服務數(LISN)
8	取自超管理器呼叫參數之中斷向量表
9	狀態暫存器(SR)值
10	邏輯分割ID(LPID)
11	真實位址(RA)超管理器加速器利用記錄指標
12	儲存描述符暫存器(SDR)

【 0158 】 超 管 理 器 將 複 數 個 加 速 器 整 合 片 490 暫 存 器 445 初 始 化 。

【 0159 】 如 圖 4F 中 所 示 ， 本 發 明 的 一 個 可 選 的 實 施 方 式 利 用 一 種 經 由 用 以 存 取 實 體 處 理 器 記 憶 體 401-402 及 GPU 記 憶 體 420-423 的 共 同 虛 擬 記 憶 體 位 址 空 間 可 定 址 的 統 一 記 憶 體 。 在 此 實 施 方 式 中 ， 執 行 於 GPU 410-413 上 之 操 作 利 用 相 同 的 虛 擬 / 有 效 記 憶 體 位 址 空 間 以 來 存 取 處 理 器 記 憶 體 401-402 且 反 之 亦 然 ， 藉 此 簡 化 可 編 程 性 。 虛 擬 / 有 效 位 址 空 間 的 第 一 部 分 被 配 置 給 處 理 器 記 憶 體 401 、 第 二 部 分 被 配 置 給 第 二 處 理 器 記 憶 體 402 、 第 三 部 分 被 配 置 給 GPU 記 憶 體 420 ， 依 此 類 推 。 整 個 虛 擬 / 有 效 記 憶 體 空 間 (有 時 稱 為 有 效 位 址 空 間) 被 藉 此 橫 跨 處 理 器 記 憶 體 401-402

及 GPU 記憶體 420-423 之各者來分配，以容許任何處理器或 GPU 利用映射至該記憶體之虛擬位址來存取任何實體記憶體。

【0160】可提供 MMU 439A-439E 中的一或多個之內的偏移/同調管理電路 494A-494E 確保了介於主機處理器 (例如，405) 與 GPU 410-413 的快取之間的快取同調，並且實現了指示某些類型的資料應被儲存之實體記憶體的偏移技術。雖然偏移/同調管理電路 494A-494E 的多個實例被繪示在圖 4F 中，但偏移/同調電路還可被實現在一或多個主機處理器 405 的 MMU 內及/或加速器積體電路 436 內。

【0161】GPU 附接的記憶體 420-423 被映射為系統記憶體的一部分，並使用共享虛擬記憶體 (SVM) 技術來存取，但不會遭受與全系統快取同調性相關的典型效能缺點。GPU 附接的記憶體 420-423 被存取為系統記憶體而沒有繁重的快取同調性負擔之能力提供了對於 GPU 卸載的有利操作環境。此配置容許主機處理器 405 軟體設定操作元並存取計算結果，而沒有傳統 I/O DMA 資料副本的負擔。這樣的傳統副本涉及驅動器呼叫、中斷及記憶體映射的 I/O (MMIO) 存取，其相對於簡單記憶體存取皆為無效率的。同時，存取 GPU 附接的記憶體 420-423 而沒有快取同調性負擔之能力可能對於卸載計算之執行時間是關鍵的。在具有實質串流寫入記憶體流量的情況下，例如，快取同調性負擔可能會顯著地減少由 GPU 410-413 所見的有效寫入頻寬。操作元設定的效率、結果存取的效率及 GPU 計算

的效率皆參與判定 GPU 卸載之有效性。

【0162】GPU 偏移與主機處理器偏移之間的選擇係由偏移追蹤器資料結構來驅動。偏移表可被使用，例如，其可以是頁面粒度的結構(亦即，以記憶體頁面之粒度來控制)，其每個 GPU 附接的記憶體頁面包括 1 或 2 位元。偏移表可被實現於一或多個 GPU 附接的記憶體 420-423 的竊取記憶體範圍中，其在 GPU 410-413 中有或沒有偏移快取(例如，用以快取偏移表中的頻繁/最近使用的項目)。替代地，整個偏移表可被保持於 GPU 內。

【0163】在一實施方式中，與 GPU 附接的記憶體 420-423 之各存取相關的偏移表項目係在對於該 GPU 記憶體的實際存取前被存取，造成以下操作。首先，來自在 GPU 偏移中發現其頁面的 GPU 410-413 的本地請求被直接地傳遞至相應的 GPU 記憶體 420-423。來自在主機偏移中發現其頁面之 GPU 的本地請求被傳遞至處理器 405(例如，透過如以上所討論的高速鏈路)。可選地，來自在主處理器偏移中發現該請求頁面之處理器 405 的請求完成了該請求似的一般記憶體讀取。可替代地，指向 GPU 偏移頁面之請求可被傳遞至 GPU 410-413。假如 GPU 目前並非正在使用該頁面，則 GPU 可接著將該頁面轉換至主處理器偏移。

【0164】頁面之偏移狀態可藉由以下中的任一者來改變：軟體為基的機制、硬體協助之軟體為基的機制或(針對有限組的情況)純粹硬體為基的機制。

【0165】一種用於改變偏移狀態之機制利用了 API 呼

叫(例如，OpenCL)，其接著呼叫GPU之裝置驅動器，其接著傳送訊息(或將一命令描述符放入佇列)至GPU以指引其改變偏移狀態，並且(針對一些轉換)在該主機中執行快取清除操作。快取清除操作對於從主機處理器405偏移到GPU偏移之轉換是必要的，但對於相反轉換並非必要的。

【0166】快取同調性係藉由主機處理器405暫時地使GPU偏移的頁面無法快取來維持。為了存取這些頁面，處理器405可從GPU 410請求存取，取決於實現該GPU 410可以或可以不立刻准予存取。因此，為了減少介於處理器405與GPU 410之間的通訊，有利的是確保GPU偏移的頁面為那些由GPU所需的，而非主機處理器405所需的，且反之亦然。

圖形處理管線

【0167】圖5繪示圖形處理管線500。圖形多處理器(諸如，如圖2D中的圖形多處理器234、圖3A中的圖形多處理器325、圖3B中的圖形多處理器350)可實施所示之圖形處理管線500。圖形多處理器可被包括在如本文中所述的並行處理子系統內，諸如圖2A的並行處理器200，其可以與圖1的並行處理器112有關，並且可以用來代替其中之一。如本文所述，各種並行處理系統可經由並行處理單元(例如，圖2A的並行處理單元202)的一個或多個實例來實施圖形處理管線500。例如，著色器單元(例如，圖2C的圖形多處理器234)可組態成執行頂點處理單元504、鑲嵌控

制處理單元 508、鑲嵌評估處理單元 512、幾何處理單元 516 及片段/像素處理單元 524 中的一或多者的功能。資料組譯器 502、基元組譯器 506、514、518、鑲嵌單元 510、柵格化器 522 及柵格操作單元 526 之功能還可由處理叢集(例如，圖 2A 的處理叢集 214)及相應分割單元(例如，圖 2A 的分割單元 220A-220N)內的其他處理引擎來執行。圖形處理管線 500 還可使用針對一或多個功能的專用處理單元來實現。圖形處理管線 500 中的一或多個部分可由通用處理器(例如，CPU)內的並行處理邏輯來執行。可選地，圖形處理管線 500 中的一或多個部分可經由記憶體介面 528(其可以是圖 2A 的記憶體介面 218 的實例)來存取晶載記憶體(例如，圖 2A 的並行處理器記憶體 222)。圖形處理管線 500 亦可經由圖 3C 中的多核心組 365A 實現。

【0168】資料組譯器 502 為一種收集表面及基元之頂點資料的處理單元。資料組譯器 502 接著將頂點資料(包括頂點屬性)輸出至頂點處理單元 504。頂點處理單元 504 為執行頂點著色器程式、照亮及轉換如由頂點著色器程式所指定的頂點資料的可編程執行單元。頂點處理單元 504 係讀取被儲存於快取、本地或系統記憶體中之資料以用於處理頂點資料，並可被編程以將頂點資料從物件為基的座標表示轉換為世界空間座標空間或正規化裝置座標空間。

【0169】基元組譯器 506 的第一實例係接收來自頂點處理單元 504 之頂點屬性。基元組譯器 506 係根據需要讀取已儲存的頂點屬性，並建構圖形基元以供鑲嵌控制處理單

元 508 處理。圖形基元包括三角形、線段、點、補丁(等等)，如由各種圖形處理應用程式編程介面(API)所支持的。

【0170】鑲嵌控制處理單元 508 將輸入頂點視為幾何補丁之控制點。控制點係從來自該補丁(例如，該補丁之基礎)的輸入表示被轉換為適於藉由鑲嵌評估處理單元 512 之表面評估的表示。鑲嵌控制處理單元 508 還可計算幾何補丁之邊緣的鑲嵌因數。鑲嵌因數適用於單一邊緣並量化與該邊緣相關細節的觀點相依程度。鑲嵌單元 510 被組態成接收補丁之邊緣的鑲嵌因數，並將該補丁鑲嵌入多個幾何基元(諸如線、三角形或四邊形基元)，其被傳輸至鑲嵌評估處理單元 512。鑲嵌評估處理單元 512 操作於子分區補丁之參數化座標上，以產生用於與幾何基元相關的各頂點之表面表示及頂點屬性。

【0171】基元組譯器 514 的第二實例接收來自鑲嵌評估處理單元 512 的頂點屬性、根據需要讀取已儲存的頂點屬性及建構圖形基元以供幾何處理單元 516 處理。幾何處理單元 516 為執行幾何著色器程式以轉換如由幾何著色器程式所指明的接收自基元組譯器 514 之圖形基元的可編程執行單元。幾何處理單元 516 被編程以將圖形基元子分割為一或多個新的圖形基元並計算被用來將新的圖形基元柵格化之參數。

【0172】幾何處理單元 516 可在幾何串流中添加或刪除元件。幾何處理單元 516 將該些參數及指明新圖形基元

之頂點輸出至基元組譯器 518。基元組譯器 518 從幾何處理單元 516 接收該些參數及頂點並建構圖形基元以供視埠縮放、剔除及修剪單元 520 處理。幾何處理單元 516 讀取其被儲存於並行處理器記憶體或系統記憶體中之資料以用於處理該幾何資料。視埠縮放、剔除及修剪單元 520 係執行修剪、剔除及視埠縮放並將處理過的圖形基元輸出至柵格化器 522。

【0173】柵格化器 522 可執行深度剔除及其他深度為基的優化。柵格化器 522 也在新的圖形基元上執行掃描轉換，以產生片段並且將那些片段及相關的覆蓋資料輸出至片段/像素處理單元 524。片段/像素處理單元 524 為一種可編程執行單元，其被組態成執行片段著色器程式或像素著色器程式。片段/像素處理單元 524 轉換如由片段或像素著色器程式所指明的接收自柵格化器 522 之片段或像素。例如，片段/像素處理單元 524 可被編程以執行操作，包括但不限於紋理映射、著色、混合、紋理校正及個別校正，以產生其被輸出至柵格操作單元 526 之已著色片段或像素。片段/像素處理單元 524 可讀取被儲存於並行處理器記憶體或系統記憶體中之資料以供處理該片段資料時使用。根據對於該些處理單元組態的取樣率，片段或像素著色器程式可被組態成在樣本、像素、磚片或其他粒度上著色。

【0174】柵格操作單元 526 為一種處理單元，其執行柵格操作，包括但不限於模板、z 測試、混合，等等，並將像素資料輸出為已處理的圖形資料以被儲存於圖形記憶

體中(例如，如圖2A中的並行處理器記憶體222，及/或如圖1中的系統記憶體104)、以被繪示在一或多個顯示裝置110上或者以利一或多個處理器102或並行處理器112之一者進一步處理。柵格操作單元526被組態成壓縮被寫入至記憶體之 z 或顏色資料，並且解壓縮被讀取自記憶體之 z 或顏色資料。

機器學習概述

【0175】上述的架構可應用於使用機器學習模型執行訓練及推理操作。機器學習已成功解決了許多任務。當訓練和使用機器學習算法(例如，類神經網路)時出現的計算自然會適合於高效的並行實施方式。因此，諸如通用圖形處理單元(GPGPU)之類的並行處理器在深度類神經網路的實際實施方式中發揮了重要作用。具有單指令多執行緒(SIMT)架構的並行圖形處理器旨在最大化圖形管線中的並行處理量。在SIMT架構中，並行執行緒組嘗試盡可能經常同步執行程式指令以提高處理效率。並行機器學習演算法所提供的效率允許使用大容量網路，並使這些網路可以在更大的資料集上進行訓練。

【0176】機器學習演算法為一種可基於一組資料而學習的演算法。例如，機器學習演算法可以被設計為在一個資料組內建模式高階抽象。例如，可以使用影像識別演算法來判定給定的輸入屬於哪個類別；迴歸演算法可輸出給定輸入的數值；以及圖案識別演算法可以用於產生翻譯的

文本或執行文本到語音及/或語音辨識。

【0177】機器學習演算法的例示性類型是類神經網路。有許多類型的類神經網路；一種簡單類型的類神經網路為前饋網路。前饋網路可以被實現為其中節點被分層佈置的非循環圖。通常，前饋網路拓撲包括由至少一個隱藏層分開的輸入層和輸出層。隱藏層將輸入層接收到的輸入變換為可用於在輸出層中產生輸出的表示。網路節點經由邊緣完全連接到相鄰層中的節點，但每層中的節點之間沒有邊緣。在前饋網路之輸入層的節點處接收到的資料經由啟用函數被傳播(即，「前饋」)到輸出層的節點，啟用函數基於分別與連接層的每個邊緣相關聯的係數(「權重」)來計算網路中的每個連續層的節點的狀態。根據正在執行的演算法所表示的具體模式，類神經網路演算法的輸出可以採取各種形式。

【0178】在可以使用機器學習演算法來模擬特定問題之前，使用訓練資料組來訓練演算法。訓練類神經網路涉及選擇網路拓撲，使用表示由網路建模的問題之一組訓練資料，並且調整權重直到網路模型以訓練資料組的所有實例的最小誤差執行。例如，在用於類神經網路的監督學習訓練過程期間，響應於表示訓練資料組中的實例的輸入由網路產生的輸出與實例之標記為「正確」的輸出進行比較、計算表示輸出和標記的輸出之間差異的誤差訊號、以及調整與連接相關聯的權重，以在誤差訊號向後傳播通過網路層時最小化誤差。當從訓練的資料組之實例產生的每

一個輸出的誤差最小化時，網路被認為是「訓練的」。

【0179】機器學習演算法的準確度會受到用於訓練演算法的資料組之品質的顯著影響。訓練處理可能是計算密集型的，並且可能在習知通用處理器上需要大量的時間。因此，並行處理硬體被用於訓練許多類型的機器學習演算法。這對於優化類神經網路的訓練特別有用，因為在調整類神經網路中的係數中執行的計算本身適合於並行實施方式。具體來說，許多機器學習演算法和軟體應用已經適應於利用通用圖形處理裝置內的並行處理硬體。

【0180】圖6為機器學習軟體堆疊600之一般化圖。機器學習應用程式602為可被組態成使用訓練資料集來訓練類神經網路或者使用經訓練的深度類神經網路來實現機器智慧的任意邏輯。機器學習應用程式602可包括用於可被用來在部署前訓練類神經網路之類神經網路及/或專門軟體的訓練及推理功能。機器學習應用程式602可實現任何類型的機器智慧，包括但不限於影像辨識、映射和本地化、自主性導航、語音合成、醫藥成像或語言翻譯。實例機器學習應用程式602包括但不限於語音虛擬助理、影像或面部識別演算法、自主性導航及用於訓練機器學習應用程式602使用的機器學習模型的軟體工具。

【0181】機器學習應用程式602的硬體加速可經由機器學習框架604啟用。機器學習框架604可提供機器學習基元之庫。機器學習基元為常由機器學習演算法所執行的基本操作。若沒有機器學習框架604，則機器學習演算法之

開發者將被要求建立並優化其與機器學習演算法相關的主計算邏輯，然後隨著新的並行處理器被開發而再優化該計算邏輯。取而代之地，機器學習應用程式可被組態成使用由機器學習框架 604 所提供的基元來執行必要的計算。示例性基元包括張量卷積、啟動功能和池化，其為訓練卷積類神經網路 (CNN) 時所執行的計算操作。機器學習框架 604 還可提供基元以實現由許多機器學習演算法 (諸如矩陣和向量操作) 來執行的基本線性代數子程式。機器學習框架 604 的實例包括但不限於 TensorFlow、TensorRT、PyTorch、MXNet、Caffee、及其他高階機器學習框架。

【0182】機器學習框架 604 可處理接收自機器學習應用程式 602 之輸入資料並產生適當的輸入給計算框架 606。計算框架 606 可提取被提供給 GPGPU 驅動器 608 的主要指令，以致使機器學習框架 604 利用經由 GPGPU 硬體 610 之硬體加速而不需要機器學習框架 604 具有 GPGPU 硬體 610 之架構的內部知識。此外，計算框架 606 可跨多種類型及世代的 GPGPU 硬體 610 來致使機器學習框架 604 之硬體加速。示例性計算框架 606 包括 CUDA 計算框架和相關的機器學習庫，例如 CUDA 深度類神經網路 (cuDNN) 庫。機器學習軟體堆疊 600 亦可包括通訊庫或框架，以促進多 GPU 和多節點計算。

GPGPU 機器學習加速

【0183】圖 7 繪示通用圖形處理單元 700，其可為圖 2A

的並行處理器 200 或圖 1 的並行處理器 112。通用圖形處理單元 (GPGPU) 700 可被組態成為由機器學習框架提供之基元的硬體加速提供支持，以加速與訓練深度類神經網路相關聯的計算工作負載之類型的處理。額外地，GPGPU 700 可被直接地鏈路至 GPGPU 的其他實例，以建立多 GPU 叢集來增進特別是用於深類神經網路之訓練速度。亦支持基元以加速已部署類神經網路的推理操作。

【0184】GPGPU 700 包括用以致使與主處理器之連接的主介面 702。主介面 702 為 PCI Express 介面。然而，主介面還可以是供應商特定的通訊介面或通訊組織。GPGPU 700 從主機處理器接收命令並使用全域排程器 704 來將與那些命令相關的執行緒分配至一組處理叢集 706A-706H。該些處理叢集 706A-706H 共享了快取記憶體 708。快取記憶體 708 可作為用於處理叢集 706A-706H 內的快取記憶體的高階快取。所示之處理叢集 706A-706H 可相應於圖 2A 中的處理叢集 214A-214N。

【0185】GPGPU 700 包括經由一組記憶體控制器 712A-712B 而與處理叢集 706A-706H 耦接之記憶體 714A-714B。記憶體 714A-714B 可包括各個類型的記憶體裝置，包括動態隨機存取記憶體 (DRAM) 或圖形隨機存取記憶體，諸如同步圖形隨機存取記憶體 (SGRAM)，包括圖形雙資料速率 (GDDR) 記憶體。記憶體 714A-714B 也可包括 3D 堆疊記憶體，其包括但不限制於高頻寬記憶體 (HBM)。

【0186】每一處理叢集 706A-706H 包括諸如圖 2D 之圖

形多處理器 234、圖 3A 之圖形多處理器 325、圖 3B 之圖形多處理器 350 的一組圖形多處理器，或可包括圖 3C 中之多核心群 365A-365N。計算叢集之圖形多處理器包括多種類型的整數及浮點邏輯單元，其可執行在包括適於機器學習計算之精確度範圍的計算操作。例如，在處理叢集 706A-706H 之各者中的浮點單元之至少一子集可被組態成執行 16 位元或 32 位元浮點操作，而浮點單元的不同子集可被組態成執行 64 位元浮點操作。

【0187】 GPGPU 700 的多個實例可被組態成操作為計算叢集。由計算叢集所用於同步化及資料交換的通訊機制係跨越實施例而改變。例如，GPGPU 700 的多個實例係透過主介面 702 來通訊。在一實施例中，GPGPU 700 包括 I/O 集線器 709，其將 GPGPU 700 與 GPU 鏈路 710 耦接，該 GPU 鏈路 710 致使直接連接至 GPGPU 的其他實例。GPU 鏈路 710 被耦接至專用 GPU 至 GPU 橋，其致使 GPGPU 700 的多個實例之間的通訊及同步化。可選地，GPU 鏈路 710 係與高速互連耦接以將資料傳輸及接收至其他 GPGPU 或並行處理器。GPGPU 700 的多個實例被置於個別的資料處理系統中並經由網路裝置(其為經由主介面 702 可存取的)來通訊。除了主介面 702 之外(或當作其替代)，GPU 鏈路 710 可被組態成致使連接至主機處理器。

【0188】雖然 GPGPU 700 之所繪示組態可被組態成訓練類神經網路，但 GPGPU 700 之替代組態可被組態以供部署於高效能或低功率推理平台內。在推理組態中，相對於

訓練組態，GPGPU 700包括較少的處理叢集706A-706H。此外，與記憶體714A-714B相關的記憶體技術可在推理與訓練組態之間有差異。在一實施例中，GPGPU 700的推理組態可支持推理特定指令。例如，推理組態可提供一或多個8位元整數內積指令之支持，其常在已部署類神經網路的推理操作期間使用。

【0189】圖8繪示多GPU計算系統800。多GPU計算系統800可包括經由主介面開關804來耦接至多GPGPU 806A-806D的處理器802。主介面開關804為PCI Express開關裝置，其將處理器802耦接至PCI Express匯流排，處理器802可透過該PCI Express匯流排而與該組GPGPU 806A-806D通訊。多個GPGPU 806A-806D之每一者可以是圖7之GPGPU 700的實例。GPGPU 806A-806D可經由一組高速點對點GPU至GPU鏈路816來互連。高速GPU至GPU鏈路可經由專用GPU鏈路(諸如圖7中之GPU鏈路710)來連接至GPGPU 806A-806D之各者。P2P GPU鏈路816致使GPGPU 806A-806D的各者之間的直接通訊而不需要透過主介面匯流排(其連接至處理器802)來通訊。利用導引至P2P GPU鏈路之GPU至GPU流量，主介面匯流排保持可用於系統記憶體存取或者與多GPU計算系統800的其他實例通訊，例如，經由一或多個網路裝置。儘管在圖8中GPGPU 806A-806D係經由主介面開關804而連接至處理器802，但是處理器802包括對於P2P GPU鏈路816之直接支持並可直接地連接至GPGPU 806A-806D。在一實施例中，P2P GPU鏈路

816使多 GPU 計算系統 800 能夠操作為單個邏輯 GPU。

機器學習類神經網路實施方式

【0190】由本文中所述之計算架構可被組態成執行多類型的並行處理，其特別適於用於機器學習之訓練及部署類神經網路。類神經網路可被一般化成為具有圖形關係之功能的網路。如本領域中眾所周知，有用於機器學習的多種類型的類神經網路實現。如先前所述的，類神經網路的一種例示性類型為前饋網路。

【0191】類神經網路之第二例示性類型為卷積類神經網路 (CNN)。CNN 為用以處理具有已知、光柵狀拓撲之資料 (諸如，影像資料) 的一種專門前饋類神經網路。因此，CNN 常用於計算視覺和影像辨識應用，但其還可被用在其他類型的圖案辨識，諸如語音及語言處理。CNN 輸入層中的節點被組織為一組「過濾器」(由視網膜中所發現的個別欄位所驅使的特徵檢測器)，且各組過濾器之輸出被傳播至該網路之連續層中的節點。CNN 之計算包括將卷積數學操作應用於各過濾器以產生該過濾器之輸出。卷積為由兩個函數所執行用以產生兩個原始函數之一的修改版本之第三函數之特殊種類的數學操作。於卷積網路術語中，至卷積的第一函數可被稱為輸入，而第二函數可被稱為卷積內核。輸出可被稱為特徵映射。例如，至卷積層之輸入可以是界定輸入影像之各種顏色成分的多維陣列資料。卷積內核可以是參數的多維陣列，其中該些參數係由類神經網

路的訓練程序所調適。

【0192】遞迴類神經網路(RNN)為前饋類神經網路之家族，其包括介於層之間的回饋連接。RNN係藉由共享跨類神經網路之不同部分的參數資料以致使連續資料之建模。RNN之架構包括週期。該些週期代表變數的目前值對於未來時刻的其本身值的影響，因為來自RNN之輸出資料的至少一部分被使用為用以處理序列中之後續輸入的回饋。此特徵使得RNN特別有用於語言處理，因為語言資料可被組成的可變本質。

【0193】以下所述之圖形呈現範例的前饋、CNN與RNN網路，以及描述用以個別地訓練並部署那些類型的網路之各者的一般程序。將理解，關於本文中所述的任何特定實施例的這些描述是範例的而非限制的，且一般來說，所顯示之概念可被通常地應用於深度類神經網路及機器學習技術。

【0194】上述例示性類神經網路可被用來執行深度學習。深度學習為使用深度類神經網路之機器學習。用於深度學習之深度類神經網路為由多個隱藏層所組成的人工類神經網路，不同於僅包括單一隱藏層的淺類神經網路。較深的類神經網路對於訓練通常在計算上是較為密集的。然而，網路的額外隱藏層致使多步驟圖案辨識，其導致相對於淺機器學習技術的降低輸出錯誤。

【0195】用於深度學習之深度類神經網路通常包括用以執行特徵辨識之前端網路，其耦接至後端網路，該後端

網路代表一種數學模型，其可根據提供至該模型之特徵表示來執行操作(例如，物件分類、語音辨識，等等)。深度學習致使機器學習被執行而不需要針對該模型執行手動調整特徵工程。取而代之地，深度類神經網路可根據輸入資料內的統計結構或相關性來學習特徵。經學習的特徵可被提供至一種可將測得特徵映射至輸出的數學模型。由該網路所使用之數學模型一般針對待執行的特定工作而被特殊化，並且不同的模型將用以執行不同的工作。

【0196】一旦類神經網路被構成，學習模型可被應用至該網路以訓練該網路來執行特定工作。該學習模型描述如何調整該模型內的權重以減少該網路之輸出錯誤。錯誤的後向傳播為用以訓練類神經網路的常見方法。輸入向量被提交至該網路以供處理。該網路之輸出係使用損失函數來與所欲輸出進行比較，並針對輸出層中之神經元的各者來計算誤差值。誤差值被接著向後傳播，直到各神經元具有約略地表示其對於原始輸出的貢獻的相關誤差值。該網路可接著使用演算法(諸如隨機梯度下降演算法)自那些錯誤學習，以更新該類神經網路之權重。

【0197】圖 9A 及 9B 繪示例示性卷積類神經網路。圖 9A 繪示 CNN 內的各個層。如圖 9A 中所示，用以建模影像處理之例示性 CNN 可接收描述輸入影像之紅、綠及藍(RGB)成分的輸入 902。輸入 902 可由多個卷積層(例如，卷積層 904、卷積層 906)來處理。來自多個卷積層之輸出可選擇性地由一組完全連接層 908 來處理。完全連接層中之

神經元具有對於先前層中之所有啟動的完全連接，如先前前饋網路所述。來自完全連接層 908 之輸出可被用來產生來自該網路之輸出結果。完全連接層 908 內的啟動可使用取代卷積的矩陣乘法來計算。並非所有 CNN 的實現皆利用完全連接層 908。例如，在一些實現中，卷積層 906 可產生輸出給該 CNN。

【0198】卷積層被稀疏地連接，其不同於完全連接層 908 中所見之傳統類神經網路組態。傳統類神經網路層被完全連接，以致使每一輸出單元係與每一輸入單元互動。然而，如圖所示，卷積層被稀疏地連接，因為欄位之卷積的輸出被輸入(取代該欄位中之各節點的個別狀態值)至後續層之節點。與該些卷積層相關的內核係執行卷積操作，其輸出被傳送至下一層。在卷積層內所執行的維度減少為致使 CNN 縮放以供處理大影像之一種形態。

【0199】圖 9B 繪示 CNN 之卷積層內的例示性計算階段。對 CNN 之卷積層 912 的輸入可在卷積層 914 的三個階段中被處理。該三個階段可包括卷積階段 916、檢測器階段 918 和池化階段 920。卷積層 914 可接著輸出資料至連續卷積層。該網路的最終卷積層可產生輸出特徵映射資料或提供輸入至完全連接層，例如，用以產生輸入的分類值給該 CNN。

【0200】在卷積階段 916 中，並行地執行數個卷積以產生一組線性啟動。卷積階段 916 可包括仿射變換，其為可被指明為線性變換加上轉換之任何變換。仿射變換包括

旋轉、轉換、縮放及這些變換之組合。卷積階段計算函數(例如，神經元)之輸出，其被連接至輸入中之特定區，其可被判定為與該神經元相關的本地區。神經元計算介於該些神經元的權重與該些神經元所連接至之該本地輸入中的該區之間的內積。來自卷積階段 916 之輸出係界定一組線性啟動，其由卷積層 914 之連續階段來處理。

【0201】該些線性啟動可由檢測器階段 918 來處理。在檢測器階段 918 中，各線性啟動係由非線性啟動函數來處理。非線性啟動函數係增加整體網路之非線性性質而不影響卷積層之個別欄位。數種類型的非線性啟動函數可被使用。一種特別類型是整流線性單元(ReLU)，其使用界定為 $f(x) = \max(0, x)$ 之啟動函數，以致使啟動係以零為臨限值。

【0202】池化階段 920 係使用以附近輸出之摘要統計來替換卷積層 906 之輸出的池化函數。池化函數可被用來將轉換不變性引入類神經網路，以致使對於輸入之小轉換不會改變池化的輸出。對於本地轉換之不變性可能在輸入資料中之特徵的存在比該特徵之精確位置更為重要的情境中是有用的。多種類型的池化函數可被用於池化階段 920 期間，包括最大池化、平均池化及 l_2 -norm 池化。此外，一些 CNN 實施方式不包括池化階段。取而代之地，此類實施方式係以相對於先前卷積階段具有增加的跨步的額外卷積階段來代替。

【0203】來自卷積層 914 之輸出可接著由下一層 922 來處理。下一層 922 可以是額外卷積層或者完全連接層 908 之

一。例如，圖 9A 的第一卷積層 904 可輸出至第二卷積層 906，而第二卷積層可輸出至完全連接層 908 的第一層。

【0204】圖 10 繪示例示性遞迴類神經網路 1000。在遞迴類神經網路 (RNN) 中，網路的先前狀態係影響網路的目前狀態的輸出。RNN 可使用多種函數來以多種方式被建立。RNN 之使用通常圍繞著使用數學模型以根據輸入之先前序列來預測未來。例如，RNN 可被用來執行統計語言建模來根據字詞之先前序列而預測將來的字詞。所繪示的 RNN 1000 可被描述為具有接收輸入向量的輸入層 1002、用以實施遞迴函數的隱藏層 1004、用以致使先前狀態之「記憶體」的回饋機制 1005 及用以輸出結果的輸出層 1006。RNN 1000 係根據時間步驟來操作。在既定時間步驟之 RNN 的狀態係經由回饋機制 1005 根據先前時間步驟而被影響。針對既定時間步驟，隱藏層 1004 之狀態係由先前狀態以及目前時間步驟之輸入來界定。在第一時間步驟之初始輸入 (x_1) 可由隱藏層 1004 來處理。第二輸入 (x_2) 可由隱藏層 1004 使用在該初始輸入 (x_1) 之處理期間所判定的狀態資訊來處理。既定狀態可被計算為 $s_t = f(Ux_t + Ws_{t-1})$ ，其中 U 和 W 為參數矩陣。函數 f 通常為非線性，諸如雙曲線正切函數 (Tanh) 或整流函數 $f(x) = \max(0, x)$ 之變化。然而，用於隱藏層 1004 之特定數學函數可根據 RNN 1000 之特定實施方式細節而變化。

【0205】除了所述的基本 CNN 及 RNN 網路之外，對於那些網路的加速變化可被致能。一種例示性 RNN 變體為長

短期記憶體(LSTM)RNN。LSTM RNN能夠學習對於處理較長的語言序列可能是必要的之長期依存性。CNN之變化為一種卷積深信網路，其具有類似於CNN之結構且係以一種類似於深信網路之方式來訓練。深信網路(DBN)為一種生成的類神經網路，其由隨機(任意)變數的多個層所組成。DBN可使用貪心非監督學習而被層接層地訓練。DBN之經學習的權重可接著藉由判定針對該類神經網路之最佳初始組的權重被用來提供預先訓練類神經網路。在進一步實施例中，啟用用於強化學習的加速。在強化學習中，人工代理藉由與環境互動來學習。代理係組態為優化某些目標以最大化累積獎勵。

【0206】圖11繪示深度類神經網路之訓練和部署。一旦既定網路已針對一工作而被構成，則類神經網路係使用訓練資料集1102而被訓練。各種訓練框架1104已被開發來致使訓練程序之硬體加速。例如，圖6之機器學習框架604可被組態成訓練框架604。訓練框架604可連結入未訓練類神經網路1106並致使未訓練神經網使用本文中所述之並行處理資源而被訓練，以產生經訓練神經網1108。

【0207】為了開始訓練程序，初始權重可被隨機地選擇或者藉由使用深信網路來預先訓練。訓練週期接著以受監督或未監督中任一者的方式被執行。

【0208】受監督學習係一種學習方法，其中訓練被執行為中介操作，諸如當訓練資料集1102包括與針對輸入的所欲輸出配對的輸入時，或其中訓練資料集包括具有已知

輸出之輸入且類神經網路之輸出被手動地分級。該網路處理該些輸入並將所得輸出與一組預期或所欲輸出進行比較。錯誤接著透過系統回傳。訓練框架 1104 可調整控制未訓練類神經網路 1106 之權重。訓練框架 1104 可提供工具來監視未訓練類神經網路 1106 如何適當地根據已知的輸入資料以朝向一適於產生正確答案之模型收斂。隨著該網路之權重被調整以改良由類神經網路所產生的輸出，訓練程序係重複地發生。訓練程序可持續，直到類神經網路達到與經訓練神經網 1108 相關之統計上理想的準確度。經訓練類神經網路 1108 可接著被部署以實現任何數目的機器學習操作，以基於新資料 1112 的輸入產生推理結果 1114。

【0209】未監督學習為一種網路嘗試使用未標示資料來訓練其本身的學習方法。因此，針對未監督學習，訓練資料集 1102 將包括輸入資料而沒有任何相關的輸出資料。未訓練類神經網路 1106 可學習未標示輸入內的學習群集，並可判定個別輸入如何與整體資料集相關。未監督訓練可被用來產生自行組織映射，其為一種能夠執行可用來減少資料之維度的操作之經訓練神經網路 1108 的類型。未監督訓練還可被用來執行異常檢測，其容許從資料的正常型態偏離之輸入資料集中的資料點的識別。

【0210】受監督及未監督訓練之變化還可被利用。半監督學習為一種訓練資料集 1102 包括相同分佈之已標示及未標示資料的混合之技術。遞增學習為受監督學習之變化，其中輸入資料被持續地用來進一步訓練模型。遞增學

習致使經訓練類神經網路 1108 調適於新資料 1112 而不會忘記初始訓練期間之網路內所教導的知識。

【0211】無論受監督或未監督，特別是針對深度類神經網路之訓練程序可能對於單一計算節點是太計算密集的。取代使用單一計算節點，計算節點之分散式網路可被用來加速訓練程序。

【0212】圖 12A 係繪示分散式學習的方塊圖。分散式學習為一種使用多個分散式計算節點來執行類神經網路之受監督或未監督訓練的訓練模型。分散式計算節點可各包括一或多個主處理器及一或多個通用處理節點，諸如圖 7 中之高度並行通用圖形處理單元 700。如圖所繪示，分散式學習可執行模型並行性 1202、資料並行性 1204 或模型和資料並行性之組合 1206。

【0213】在模型並行性 1202 中，分散式系統中的不同計算節點可針對單一網路之不同部分來執行訓練計算。例如，類神經網路之各層可由分散式系統中的不同處理節點來訓練。模型並行性之優點包括縮放至特別大的模型之能力。將與類神經網路之不同層相關的計算分離致使了非常大的類神經網路(其中所有層之權重將不配適入單一計算節點之記憶體)之訓練。在一些實例中，模型並行性可特別有用於執行大型類神經網路之未監督訓練。

【0214】在資料並行性 1204 中，分散式網路的不同節點具有該模型的完整實例且各節點係接收資料之不同部分。來自不同節點之結果被接著結合。雖然資料並行性的

不同方式是可能的，但資料並行性訓練皆需要一種結合結果並將各節點間之模型參數同步化的技術。結合資料的例示性方案包括參數平均及更新為基的資料並行性。參數平均係訓練該訓練資料之子集上的各節點，並將全域參數(例如，權重、偏移)設為來自各節點之參數的平均。參數平均係使用保存參數資料的中央參數伺服器。更新為基的資料並行性係類似於參數平均，除了對於該模型的更新被轉移，而不是將參數從節點轉移至參數伺服器。此外，更新為基的資料並行性可用一種分散方式來執行，其中該些更新被壓縮並在節點之間轉移。

【0215】結合的模型及資料並行性1206可被實現在例如各計算節點包括多個GPU的分散式系統中。各節點可具有該模型的完全實例，其中各節點內的分離GPU被用來訓練該模型之不同部分。

【0216】相對於單一機器上之訓練，分散式訓練具有增加的負擔。然而，本文中所述之並行處理器及GPGPU可各實現多種技術來減少分散式訓練之負擔，包括用以致使高頻寬GPU至GPU資料轉移及加速遠端資料同步化的技術。

【0217】圖12B係繪示可編程網路介面1210及資料處理單元的方塊圖。可編程網路介面1210係可用來加速分散式環境內基於網路計算任務的可編程網路引擎。可編程網路介面1210可經由主介面1270與主機系統耦接。可編程網路介面1210可用於加速主機系統的CPU或GPU的網路或儲

存操作。主機系統可為(例如)用以執行分散式訓練的分散式學習系統的節點，例如，如圖 12A 所示。主機系統亦可為資料中心內的資料中心節點。

【0218】在一實施例中，可藉由可編程網路介面 1210 加速對含有模型資料的遠端儲存的存取。例如，可編程網路介面 1210 可被組態以將遠端儲存裝置作為本地儲存裝置呈現給主機系統。可編程網路介面 1210 亦可加速在主機系統的 GPU 與遠端系統的 GPU 之間執行的遠端直接記憶體存取 (RDMA) 操作。在一實施例中，可編程網路介面 1210 可以啟用儲存功能，諸如但不限於 NVME-oF。可編程網路介面 1210 亦可代表主機系統加速用於遠端儲存的加密、資料完整性、壓縮及其他操作，從而允許遠端儲存接近直接附接至主機系統的儲存裝置的等待時間。

【0219】可編程網路介面 1210 亦可代表主機系統執行資源分配和管理。可以將儲存安全操作卸載至可編程網路介面 1210，並與遠端儲存資源的分配及管理一起執行。替代地，可以由可編程網路介面 1210 來執行基於網路的操作來管理對遠端儲存的存取，否則該操作將由主機系統的處理器執行。

【0220】在一實施例中，網路及/或資料安全操作可從主機系統卸載至可編程網路介面 1210。資料中心節點的資料中心安全策略可以由可編程網路介面 1210 替代主機系統的處理器來處理。例如，可編程網路介面 1210 可檢測並減輕對主機系統之嘗試基於網路的攻擊(例如，DDoS)，從

而防止攻擊損害主機系統的可用性。

【0221】可編程網路介面 1210 可包括系統單晶片 (SoC 1220)，其經由多處理器核心 1222 執行作業系統。處理器核心 1222 可包括通用處理器 (例如，CPU) 核心。在一實施例中，處理核心 1222 亦可包括一或多個 GPU 核心。SoC 1220 可執行儲存在記憶體裝置 1240 中的指令。儲存裝置 1250 可儲存本地作業系統資料。儲存裝置 1250 及記憶體裝置 1240 亦可用以為主機系統快取遠端資料。網路埠 1260A-1260B 啟用到網路或結構的连接，並有助於 SoC 1220 以及主機系統經由主介面 1270 的網路存取。可編程網路介面 1210 亦可包括 I/O 介面 1275，諸如 USB 介面。I/O 介面 1275 可用於將外部裝置耦接至可編程網路介面 1210 或作為除錯介面。可編程網路介面 1210 亦包括管理介面 1230，其使主機裝置上的軟體能夠管理和組態可編程網路介面 1210 及 / 或 SoC 1220。在一實施例中，可編程網路介面 1210 亦包括一或多個加速器或 GPU 1245，以接受來自經由網路埠 1260A-1260B 耦合之 SoC 1220、主機系統、或遠端系統之並行計算任務的卸載。

例示性機器學習應用

【0222】機器學習可被應用來解決多種技術問題，包括但不限於電腦視覺、自動駕駛和導航、語音辨識及語言處理。電腦視覺傳統上已成為針對機器學習應用的最活躍研究領域之一。電腦視覺之應用的範圍從再生人類視覺能

力(諸如辨識人臉)至產生新種類的視覺能力。例如，電腦視覺應用可被組態成從視訊中可見的物體中所引發的振動來辨識音波。並行處理器加速機器學習係致使電腦視覺應用使用相較於先前可行者顯著地更大的訓練資料集來訓練，並致使推理系統使用低功率並行處理器來被部署。

【0223】並行處理器加速機器學習具有自動駕駛應用，包括巷道和道路標誌辨識、障礙閃避、導航及駕駛控制。加速機器學習技術可被用來根據界定對於特定訓練輸入之適當回應的資料集來訓練駕駛模型。本文中所述之並行處理器可致使用於自動駕駛解決方案的越來越複雜的類神經網路之快速訓練，並致使一種適於整合在自動車內的行動平台中的低功率推理處理器之部署。

【0224】並行處理器加速深度類神經網路已致使自動語音辨識(ASR)的機器學習方案。ASR包括產生計算既定輸入聲波序列下最可能的語言序列之函數。使用深度類神經網路之加速機器學習已致使先前用於ASR的隱藏馬可夫(Markov)模型(HMM)及高斯(Gaussian)混合模型(GMM)的替換。

【0225】並行處理器加速機器學習還可被用來加速自然語言處理。自動學習程序可利用統計推理演算法來產生能抵抗錯誤或不熟悉輸入的模型。例示性自然語言處理器應用包括人類語言之間的自動機器翻譯。

【0226】用於機器學習的並行處理平台可被劃分為訓練平台及部署平台。訓練平台通常是高度並行的，且包括

用以加速多 GPU 單節點訓練及多節點多 GPU 訓練的優化。適合訓練的例示性並行處理器包括圖 7 的通用圖形處理單元 700 及圖 8 的多 GPU 計算系統 800。反之，已部署的機器學習平台通常包括較低功率的並行處理器，其適於諸如相機、自動機器人及自動車等產品的使用。

【0227】額外地，機器學習技術可應用於加速或擴增圖形處理活動。例如，可以訓練機器學習模型以識別 GPU 加速應用程式產生的輸出並產生該輸出的升級版本。可將此技術應用於為遊戲應用程式加速高解析度影像的產生。機器學習的使用可使其他各種圖形管線活動受益。例如，可訓練機器學習模型以對幾何資料執行鑲嵌細分操作，以增加幾何模型的複雜性，從而可以從相對較低細節的幾何自動產生詳細的幾何。

【0228】圖 13 繪示適於使用訓練模型執行推理的例示性推理系統單晶片 (SOC) 1300。SOC 1300 可整合包括媒體處理器 1302、視覺處理器 1304、GPGPU 1306 及多核心處理器 1308 之處理組件。GPGPU 1306 可為本文所述之 GPGPU (諸如 GPGPU 700)，並且多核心處理器 1308 可為本文所述之多核心處理器 (諸如，多核心處理器 405-406)。SOC 1300 可額外地包括晶載記憶體 1305，其可致使可由各處理組件所存取之共享的晶載資料池。處理組件可針對低功率操作來優化，以致使對於包括自動車及自動機器人的多種機器學習平台之部署。例如，SOC 1300 的一種實現可被使用為自動車的主控制系統的一部分。當 SOC 1300 被組

態為用於自動車時，SOC被設計並組態成符合部署管轄權之相關的功能安全性標準。

【0229】在操作期間，媒體處理器 1302 及視覺處理器 1304 可配合加速電腦視覺操作來運作。媒體處理器 1302 可致使多個高解析度(例如，4K、8K)視訊串流的低延遲時間解碼。經解碼的視訊串流可被寫入至晶載記憶體 1305 中的緩衝器。視覺處理器 1304 可接著解析該經解碼視訊並對於該經解碼視訊之框執行初步處理操作，以準備使用經訓練的影像辨識模型來處理該些框。例如，視覺處理器 1304 可加速被用來對於高解析度視訊資料執行影像辨識之 CNN 的卷積操作，而後端模型計算係由 GPGPU 1306 來執行。

【0230】多核心處理器 1308 可包括用以協助由媒體處理器 1302 和視覺處理器 1304 所執行的資料轉移及共享記憶體操作之排序和同步化的控制邏輯。多核心處理器 1308 還可作用為用以執行可利用 GPGPU 1306 之推理計算能力的軟體應用程式的應用程式處理器。例如，導航和駕駛邏輯之至少一部分可被實現在多核心處理器 1308 上所執行的軟體。此軟體可直接地發送計算工作負載至 GPGPU 1306，或者該計算工作負載可被發送至多核心處理器 1308，其可將那些操作之至少一部分卸載至 GPGPU 1306。

【0231】GPGPU 1306 可包括計算叢集，諸如通用圖形處理單元 700 內的處理叢集 706A-706H 的低功率組態。GPGPU 1306 內的計算叢集可支持被明確地優化以執行推理計算於經訓練類神經網路上的指令。例如，GPGPU

1306可支持用以執行諸如8位元及4位元整數向量操作的低精確度計算的指令。

額外系統概述

【0232】圖14為根據實施例之處理系統1400的方塊圖。具有與本文中任何其他圖式之元件相同或相似名稱之圖14的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實體，但是不限於此。系統1400可用於單一處理器桌上型系統、多處理器工作站系統、或具有大量處理器1402或處理核心1407的伺服器系統中。系統1400可為整合於系統單晶片(SoC)積體電路的處理平台，以用於形成、手持或嵌入式裝置，諸如具有區域或廣域網路的有線或無線連接的物聯網(IoT)裝置。

【0233】系統1400可以是具有相應於圖1之組件之組件的處理系統。例如，在不同組態中，處理器1402或處理核心1407可對應於圖1之處理器102。圖形處理器1408可對應於圖1之並行處理器112。外部圖形處理器1418可為圖1之外接裝置120中的一者。

【0234】系統1400可包括、耦接或被結合入伺服器為基的遊戲平台、遊戲控制台，包括遊戲和媒體控制台、行動遊戲控制台、手持式遊戲控制台或線上遊戲控制台。系統1400可為行動電話、智慧型手機、平板計算裝置、或連接行動網路的裝置(諸如具有低內部儲存容量的膝上型電

腦)的一部分。處理系統 1400 亦可包括下列中的一者、與下列中的一者耦接或整合在下列中的一者內：穿戴裝置(諸如智能手錶穿戴裝置)；擴增實境(AR)或虛擬實境(VR)功能擴增的智慧眼鏡或衣服，其可提供視覺、音頻或觸覺輸出，以補充現實世界中的視覺、音頻或觸覺體驗、或者提供文字、音頻、影像、視訊，全息圖像或視訊或觸覺反饋、全像影像或視訊；其他擴增實境(AR)裝置；或其他虛擬實境(VR)裝置。處理系統 1400 可包括或可為電視或機上盒裝置的一部分。系統 1400 可包括自動駕駛車輛、與自動駕駛車輛耦接或整合在自動駕駛車輛內，諸如巴士、牽引車、汽車、摩托車或電動自行車、飛機或滑翔機(或其任意組合)。自動駕駛車輛可以使用系統 1400 來處理在車輛周圍感測到的環境。

【0235】一或多個處理器 1402 各包括用以處理指令的一或多個處理器核心 1407，當所述指令被執行時，執行針對系統或使用者軟體之操作。一或多個處理器核心 1407 中的至少一者可被組態成處理特定指令集 1409。指令集 1409 可促進複雜指令集計算(CISC)、精簡指令集計算(RISC)或經由超長指令字組(VLIW)之計算。一或多個處理器核心 1407 可處理不同的指令集 1409，其可包括用以協助其他指令集之仿真的指令。處理器核心 1407 還可包括其他處理裝置，諸如數位訊號處理器(DSP)。

【0236】處理器 1402 包括快取記憶體 1404。根據該架構，處理器 1402 可以具有單一內部快取或多階內部快取。

在一些實施例中，快取記憶體被共享於處理器 1402 的各個組件之間。在一些實施例中，處理器 1402 亦使用外部快取(例如，例如，第 3 階 (L3) 快取或最末階快取 (LLC))(未圖示)，其可使用已知的快取同調性技術而被共享於處理器核心 1407 之間。暫存器檔案 1406 可被額外地包括在處理器 1402 中，並且可包括不同類型的暫存器，用以儲存不同類型的資料(例如，整數暫存器、浮點暫存器、狀態暫存器及指令指標暫存器)。一些暫存器可以是通用暫存器，而其他暫存器可特別針對處理器 1402 之設計。

【0237】一或多個處理器 1402 可與一或多個介面匯流排 1410 耦接，用以在處理器 1402 與系統 1400 中的其他組件之間傳輸通訊信號(諸如位址、資料或控制信號)。在這些實施例之一中，介面匯流排 1410 可以為處理器匯流排，諸如直接媒體介面 (DMI) 匯流排的版本。然而，處理器匯流排不限於 DMI 匯流排，並且可包括一或多個週邊組件互連匯流排(例如，PCI、PCI express)、記憶體匯流排、或其他類型介面匯流排。例如，處理器 1402 可包括積體記憶體控制器 1416 及平台控制器集線器 1430。記憶體控制器 1416 促進記憶體裝置與系統 1400 之其他組件之間的通訊，而平台控制器集線器 (PCH) 1430 經由本地 I/O 匯流排提供連接至 I/O 裝置。

【0238】記憶體裝置 1420 可以是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、相變記憶體裝置或具有用以作用為程序記憶體

之適當效能的一些其他記憶體裝置。記憶體裝置 1420 可(例如)可操作為用於系統 1400 之系統記憶體，用以儲存資料 1422 及指令 1421，以便在當一或多個處理器 1402 執行應用程式或程序時使用。記憶體控制器 1416 亦與選擇性的外部圖形處理器 1418 耦接，該外部圖形處理器 1418 可與處理器 1402 中的一或多個圖形處理器 1408 通訊以執行圖形及媒體操作。在一些實施例中，圖形、媒體、及/或計算操作可由加速器 1412 輔助，加速器 1412 係可組態以執行一組專用的圖形、媒體或計算操作的協同處理器。例如，加速器 1412 可以用於優化機器學習或計算操作的矩陣乘法加速器。加速器 1412 可以是射線追蹤加速器，其可以用於與圖形處理器 1408 協同執行射線追蹤操作。在一實施例中，可以使用外部加速器 1419 取代加速器 1412 或與之配合使用。

【0239】可以提供可以連接到處理器 1402 的顯示裝置 1411。顯示裝置 1411 可以是內部顯示裝置中的一或多個，例如在行動電子裝置或膝上型裝置中，或者是經由顯示介面(例如，DisplayPort 等)附接的外部顯示裝置。顯示裝置 1411 可為頭戴式顯示器(HMD)，諸如用於虛擬實境(VR)應用程序或擴增實境(AR)應用程序中的立體顯示裝置。

【0240】平台控制器集線器 1430 可以使週邊設備經由高速 I/O 匯流排連接至記憶體裝置 1420 和處理器 1402。I/O 週邊設備包括但不限於音頻控制器 1446、網路控制器 1434、韌體介面 1428、無線收發器 1426、觸控感測器 1425、資料儲存裝置 1424(例如，非揮發性記憶體、揮發

性記憶體、硬碟機、快閃記憶體、NAND、3D NAND、3D XPoint/Optane等)。資料儲存裝置1424可經由儲存介面(例如，SATA)或經由週邊匯流排(諸如，週邊組件互連匯流排(例如，PCI、PCI express))連接。觸控感測器1425可以包括觸控螢幕感測器、壓力感測器或指紋感測器。無線收發器1426可以是Wi-Fi收發器、藍牙收發器或諸如3G、4G、5G或長期演進(LTE)收發器的行動網路收發器。韌體介面1428使得能與系統韌體的通訊，並且可為(例如)統一的可擴展韌體介面(UEFI)。網路控制器1434可以啟用至有線網路的網路連接。在一些實施例中，高效能網路控制器(未圖示)係與介面匯流排1410耦接。音頻控制器1446可以是多通道高清晰度音頻控制器。在一些這些實施例中，系統1400包括可選地舊式I/O控制器1440，用於將舊式(例如，個人系統2(PS/2))裝置耦接至系統。平台控制器集線器1430亦可連接到一或多個通用串列匯流排(USB)控制器1442連接輸入裝置，諸如鍵盤和滑鼠1443的組合、相機1444、或其他USB輸入裝置。

【0241】將理解，所示之系統1400為範例性而非限制性的，因為被不同地組態的其他類型的資料處理系統也可被使用。例如，記憶體控制器1416及平台控制器集線器1430的實例可以整合在分立外部圖形處理器，諸如外部圖形處理器1418。平台控制器集線器1430及/或記憶體控制器1416可以在一或多個處理器1402的外部。例如，系統1400可包括外部記憶體控制器1416及平台控制器集線器

1430，其可組態為與處理器 1402 通訊的系統晶片組內的記憶體控制器集線器和週邊控制器集線器。

【0242】例如，其上放置諸如 CPU、記憶體和其他組件的電路板(「滑槽」)被設計用於增加熱效能。諸如處理器的處理組件可位於滑槽的頂部，而靠近的記憶體(諸如 DIMM)位於滑槽的底部。由於這種設計所提供的擴增的氣流，這些組件可以在比典型系統更高的頻率和電力水平下工作，從而提高了效能。此外，滑槽被組態為盲目地與框架中的電力和資料通訊電纜配合，從而擴增其快速移除、升級、重新安裝及/或更換的能力。類似地，位於滑槽上的單獨的組件(諸如，處理器、加速器、記憶體和資料儲存器驅動)由於它們彼此之間的間距而被組態以簡單地升級。於所繪示的實施例中，組件額外地包括硬體認證特徵以證明其真偽。

【0243】資料中心可利用支持包括乙太網路和全向路徑的多種其他網路架構的單個網路架構(「結構」)。滑槽可經由光纖耦接到開關，所述光纖提供比典型的雙絞線電纜(例如，類別 5、類別 5e、類別 6 等)更高的頻寬和更低的延遲。由於高頻寬、低潛時的互連和網路架構，資料中心可以在使用中匯集資源(諸如記憶體、加速器(例如圖形加速器、FPGA、ASIC、類神經網路及/或人工智慧加速器等))及實體分離的資料儲存驅動，並且根據需要提供它們來計算資源(例如，處理器)，使計算資源能夠存取匯集的資源，就好像它們是本地的一樣。

【0244】電源供應或電源可向系統1400或本文敘述之任何組件或系統提供電壓及/或電流。在一個實例中，電源供應包括插入牆壁插座之AC至DC(交流電流至直流電流)配接器。這樣的AC電源可以是可再生能源(例如，太陽能)電源。在一個實例中，電源包括DC電源，諸如外部AC至DC轉換器。電源供應或電源亦可包括無線充電硬體，以經由靠近充電場進行充電。電源可包括內部電池、交流電源、基於運動的電源、太陽能電源或燃料電池電源。

【0245】圖15A-15C繪示計算系統及圖形處理器。具有與本文中任何其他圖式之元件相同或相似名稱之圖15A-15C的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實體，但是不限於此。

【0246】圖15A係處理器1500的方塊圖，處理器1500可為處理器1402的一變體，並且可以用來代替那些處理器之一。因此，本文結合處理器1500的任何特徵之揭露亦揭露了與處理器1402的相應組合，但不限於此。處理器1500具有一或多個處理器核心1502A-1502N、積體記憶體控制器1514和積體圖形處理器1508。在排除積體圖形處理器1508的情況下，包括該處理器的系統將包括系統晶片組內或經由系統匯流排耦接的圖形處理器裝置。處理器1500可包括額外核心高達(並包括)由虛線方盒所表示的額外核心1502N。處理器核心1502A-1502N之各者包括一或多個內部快取單元1504A-1504N。在一些實施例中，各處理器核

心 1502A-1502N 亦得以存取一或多個共享快取單元 1506。內部快取單元 1504A-1504N 及共享快取單元 1506 係代表處理器 1500 內的快取記憶體階層。快取記憶體階層在各處理器核心內可包括至少一階的指令和資料快取，以及一或多階的共享中階快取，諸如第 2 階 (L2)、第 3 階 (L3)、第 4 階 (L4) 或其他階快取，其中在外部記憶體前的最高階快取被歸類為 LLC。在一些實施例中，快取同調性邏輯係維持介在各個快取單元 1506 及 1504A-1504N 之間的同調性。

【0247】處理器 1500 亦可包括一組一或多個匯流排控制器單元 1516 及系統代理核心 1510。一或多個匯流排控制器單元 1516 管理一組週邊裝置匯流排，諸如一或多個 PCI 或 PCI Express 匯流排。系統代理核心 1510 針對各個處理器組件提供管理功能。系統代理核心 1510 包括一或多個積體記憶體控制器 1514，用以管理對各種外部記憶體裝置 (未圖示) 之存取。

【0248】例如，一或多個處理器核心 1502A-1502N 可包括支持同時多執行緒。系統代理核心 1510 包括用以在多執行緒處理期間協調並操作核心 1502A-1502N 之組件。系統代理核心 1510 可額外地包括電力控制單元 (PCU)，其包括用以調節處理器核心 1502A-1502N 及圖形處理器 1508 之電力狀態的邏輯和組件。

【0249】處理器 1500 額外地包括用以執行圖形處理操作之圖形處理器 1508。在一些這些實施例中，圖形處理器 1508 與該組共享快取單元 1506 及系統代理核心 1510，包括

一或多個積體記憶體控制器 1514 耦接。系統代理核心 1510 亦可包括顯示控制器 1511，其驅動圖形處理器輸出至一或多個耦接的顯示器。顯示控制器 1511 可以是個別的模組，其經由至少一互連而與圖形處理器耦接，或者可被整合於圖形處理器 1508 內。

【0250】環狀互連單元 1512 可被用來耦接處理器 1500 之內部組件。然而，可使用替代的互連單元，諸如點對點互連、切換式互連或其他技術，包括本領域中眾所周知的技術。在一些關於環狀互連 1512 的這些實施例中，圖形處理器 1508 係經由 I/O 鏈路 1513 而與環狀互連 1512 耦接。

【0251】例示性 I/O 鏈路 1513 代表多種 I/O 互連中的至少一者，包括封裝上 I/O 互連，其協助各個處理器組件與高效能嵌入式記憶體模組 1518 (諸如，eDRAM 模組) 之間的通訊。可選地，處理器核心 1502A-1502N 及圖形處理器 1508 之各者係使用嵌入式記憶體模組 1518 為共享的最末階快取。

【0252】處理器核心 1502A-1502N 可為 (例如) 執行相同指令集架構之同質核心。替代地，處理器核心 1502A-1502N 就指令集架構 (ISA) 而言為異質的，其中處理器核心 1502A-1502N 中的一或多者係執行第一指令集，而其他核心中的至少一者係執行該第一指令集之子集或不同的指令集。處理器核心 1502A-1502N 就微架構而言為異質的，其中具有相對較高功率消耗中的一或多個核心係與具有較低功率消耗中的一或多個電力核心耦接。作為另一實例，處

理器核心 1502A-1502N 就計算能力而言為異質的。此外，處理器 1500 可被實現於一或多個晶片上，或者作為具有所繪示的組件以及其他組件之 SoC 積體電路。

【0253】圖 15B 係根據本文敘述的一些實施例之圖形處理器核心 1519 的硬體邏輯的方塊圖。圖形處理器核心 1519 (有時稱為核心片) 可為在模塊化圖形處理器一或多個多圖形核心。圖形處理器核心 1519 是一個圖形核心切片的實例，並且本文所述的圖形處理器可以包括基於目標電源和效能包絡線的多個圖形核心切片。每一圖形處理器核心 1519 可包括與多個子核心 1521A-1521F 耦接的固定功能塊 1530，子核心 1521A-1521F 亦稱為子片，其包括通用和固定功能邏輯的模組塊。

【0254】固定功能塊 1530 可包括例如可以在較低效能及/或較低功率圖形處理器實施方式中由圖形處理器核心 1519 中的所有子核心共享的幾何/固定功能管線 1531。幾何/固定功能管線 1531 可以包括 3D 固定功能管線 (例如，如下所述之圖 16A 中的 3D 管線 1612)、視訊前端單元、執行緒生產器 (thread spawner) 和執行緒分派器 (thread dispatcher)、及統一返回緩衝器管理器，其管理統一返回緩衝器 (例如，如下所述，圖 17 中的統一返回緩衝器 1718)。

【0255】固定功能塊 1530 亦可包括圖形 SoC 介面 1532、圖形微控制器 1533、及媒體管線 1534。圖形 SoC 介面 1532 在晶片積體電路上的系統內的圖形處理器核心 1519

和其他處理器核心之間提供介面。圖形微控制器 1533 係可編程子處理器，其可配置為管理圖形處理器核心 1519 的各種功能，包括執行緒分配、排程、及先佔(pre-emption)。媒體管線 1534(例如，圖 16A 和圖 17 的媒體管線 1616)包括有助於對包括影像、視訊資料的多媒體資料資料進行解碼、編碼、預處理及/或後處理的邏輯。媒體管線 1534 經由對子核心 1521-1521F 內的計算或採樣邏輯的請求來實施媒體操作。

【0256】SoC 介面 1532 可使圖形處理器核心 1519 能夠與通用應用程序處理器核心(例如，CPU)及/或其他 SoC 內的其他組件通訊，包括記憶體層級元件，諸如共享最後一級快取記憶體、系統 RAM 及/或嵌入式晶載或封裝上 DRAM。SoC 介面 1532 亦可使能夠與 SoC 內的固定功能裝置(例如，相機影像管線)通訊，並且使能夠使用及/或實施可在圖形處理器核心 1519 和 SoC 內 CPU 之間共享的全域記憶體原子。SoC 介面 1532 亦可為圖形處理器核心 1519 實施電源管理控制，並啟用圖形核心 1519 的時脈域與 SoC 內的其他時脈域之間的介面。可選地，SoC 介面 1532 使得能夠從命令串流器及全域執行緒分派器接收命令緩衝器，所述命令串流器及全域執行緒分派器係組態以向圖形處理器內的一或多個圖形核心中的每一者提供命令和指令。當要執行媒體操作時可以將命令和指令分派至媒體管線 1534，或者當圖形處理操作待執行時將命令和指令分派至幾何和固定功能管線(例如，幾何和固定功能管線 1531、幾何和固定

功能管線 1537)。

【0257】圖形微控制器 1533 可組態以對圖形處理器核心 1519 執行各種排程和管理任務。在一種組態中，圖形微控制器 1533 可(例如)在子核心 1521A-1521F 內的執行單元(EU)陣列 1522A-1522F、1524A-1524F 內的各種圖形並行引擎上執行圖形及/或工作負載排程。在此工作負載排程中，在包括圖形處理器核心 1519 的 SoC 的 CPU 核心上執行的主機軟體可以將工作負載提交給多個圖形處理器門鈴之一者，其在合適的圖形引擎上叫用排程操作。排持操作包括判定接下來要運行的工作負載、將工作負載提交給命令串流器、先佔引擎上正在運行的現有工作負載、監視工作負載的進度、以及在工作負載完成時通知主機軟體。可選地，圖形微控制器 1533 亦可獨立於系統上的作業系統及/或圖形驅動器軟體，促使圖形處理器核心 1519 的低功率或空閒狀態、為圖形處理器核心 1519 提供跨低功率狀態轉換保存和恢復圖形處理器核心 1519 中的暫存器的能力。

【0258】圖形處理器核心 1519 可具有比所示的子核心 1521A-1521F 多於或少於 N 個模塊化子核心。對於每組 N 個子核心，圖形處理器核心 1519 亦可包括共享功能邏輯 1535、共享及/或快取記憶體 1536、幾何/固定功能管線 1537、以及用以加速各種圖形和計算處理操作之額外的固定功能邏輯 1538。共享功能邏輯 1535 可包括與圖 17 之共享功能邏輯 1720 相關聯的邏輯單元(例如，取樣器、數學及/或執行緒間通訊邏輯)，其可由圖形處理器核心 1519 內的

每一 N 個子核心共享。共享及/或快取記憶體 1536 可為圖形處理器核心 1519 內的 N 個子核心 1521A-1521F 的最後一級快取，並且還可以用作可由多個子核心存取的共享記憶體。可在固定功能塊 1530 內包括幾何/固定功能管線 1537，而不是幾何/固定功能管線 1531，並且可包括相同或相似的邏輯單元。

【0259】圖形處理器核心 1519 可包括額外的固定功能邏輯 1538，其可包括供圖形處理器核心 1519 使用的各種固定功能加速邏輯。可選地，額外的固定功能邏輯 1538 包括用於唯位置著色 (position only shading) 的額外的幾何管線。在唯位置著色中，存在兩個幾何管線，幾何/固定功能管線 1538、1531 內的完整幾何管線以及揀選管線，其是可以包含在額外的固定功能邏輯 1538 中的額外幾何管線。例如，揀選管線可為完整幾何管線的修整版。完整管線和揀選管線可執行相同應用程式的不同實例，每一實例具有分離上下文。唯位置著色可以隱藏長時間揀選的三角形，從而著色在一些實例下可更早完成。例如，額外的固定功能邏輯 1538 內的揀選管線邏輯可以與主應用程式並行執行位置著色器，並且通常比完整管線更快地產生關鍵結果，因為揀選管線僅提取和著色頂點的位置屬性，而不執行光柵化和將像素渲染到訊框緩衝器。揀選管線可使用產生的關鍵結果來計算所有三角形的可見性資訊，而無需考慮這些三角形是否被揀選。完整管線 (在此實例中可成為重播管線) 可使用可見性資訊來跳過揀選的三角形，以便僅對

最終傳遞到柵格化階段的可見三角形進行著色。

【0260】可選地，額外的固定功能邏輯 1538 亦可包括機器學習加速邏輯，諸如固定功能矩陣乘法邏輯，用於包括針對機器學習訓練或推理之優化的實施方式。

【0261】在每一圖形子核心 1521A-1521F 中，包括一組執行資源，其可用於回應於圖形管線、媒體管線或著色器程式的請求來執行圖形、媒體及計算操作。圖形子核心 1521A-1521F 包括多個 EU 陣列 1522A-1522F、1524A-1524F、執行緒分派和執行緒間通訊 (TD/IC) 邏輯 1523A-1523F、3D(例如紋理)取樣器 1525A-1525F、媒體取樣器 1526A-1526F、著色器處理器 1527A-1527F 和共享本地記憶體 (SLM) 1528A-1528F。EU 陣列 1522A-1522F、1524A-1524F 每一個都包含多個執行單元，其為能夠為圖形、媒體或計算操作執行浮點和整數/固定點邏輯操作(包括圖形、媒體或計算著色器程式)的通用圖形處理單元。TD/IC 邏輯 1523A-1523F 對子核心內的執行單元執行本地執行緒分配和執行緒控制操作，並促進在子核心的執行單元上執行的執行緒之間的通訊。3D 取樣器 1525A-1525F 可以將與紋理或其他 3D 圖形相關的資料讀取到記憶體中。3D 取樣器可基於組態的採樣狀態和與給定紋理相關聯的紋理格式不同地讀取紋理資料。媒體取樣器 1526A-1526F 可基於與媒體資料相關聯的類型和格式來執行相似的讀取操作。例如，每一圖形子核心 1521A-1521F 可替代地包括統一 3D 及媒體取樣器。每一子核心 1521A-1521F 內執行單元上執行

的執行緒可利用每一子核心內的共享本地記憶體 1528A-1528F，以使在執行緒群組內執行的執行緒能夠使用晶載記憶體之公共池執行。

【0262】圖 15C 係根據本文敘述的實施例之可組態為圖形處理器(例如圖形處理器 1508)及/或計算加速器的通用圖形處理單元(GPGPU)1570的方塊圖。GPGPU 1570可經由一或多個系統及/或記憶體匯流排與主機處理器(例如，一或多個 CPU 1546)及記憶體 1571、1572互連。記憶體 1571可為與一或多個 CPU 1546共享的系統記憶體，而記憶體 1572係專用於 GPGPU 1570的裝置記憶體。例如，GPGPU 1570和裝置記憶體 1572內的組件可以映射至一或多個 CPU 1546可存取的記憶體位址。可經由記憶體控制器 1568促進對記憶體 1571和 1572的存取。記憶體控制器 1568可包括內部直接記憶體存取(DMA)控制器 1569，或可包括用以執行操作否則將由 DMA 控制器執行的邏輯。

【0263】GPGPU 1570包括多個快取記憶體，其包括 L2快取 1553、L1快取 1554、指令快取 1555、及共享記憶體 1556，它們中的至少一部分亦可劃分為快取記憶體。GPGPU 1570亦包括多個計算單元 1560A-1560N。每一計算單元 1560A-1560N包括一組向量暫存器 1561、純量暫存器 1562、向量邏輯單元 1563、及純量邏輯單元 1564。計算單元 1560A-1560N亦可包括本地共享記憶體 1565及程式計數器 1566。計算單元 1560A-1560N可與恆定快取 1567耦接，恆定快取 1567可用於儲存恆定資料，恆定資料是在 GPGPU

1570上執行的內核或著色器程式的運行期間不會改變的資料。恆定快取1567可為純量資料快取，並且可將快取的資料直接的提取至純量暫存器1562中。

【0264】在操作期間，一或多個CPU 1546可將命令寫入已映射至可存取位址空間的GPGPU 1570中的暫存器或記憶體中。命令處理器1557可以從暫存器或記憶體中讀取命令，並判定如何在GPGPU 1570中處理這些命令。然後可使用執行緒分派器1558將執行緒分派給計算單元1560A-1560N以執行那些命令。每一計算單元1560A-1560N可獨立於其他計算單元執行執行緒。額外地，每一計算單元1560A-1560N可獨立地組態用於條件計算，並且可將計算結果有條件地輸出至記憶體。當完成提交的命令時，命令處理器1557可以中斷一或多個CPU 1546。

【0265】圖16A-16C繪示由本文所述的實施例提供的(例如，根據圖15A-15C)額外的圖形處理器和計算加速器架構的方塊圖。具有與本文中任何其他圖式之元件相同或相似名稱之圖16A-16C的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實體，但是不限於此。

【0266】圖16A係圖形處理器1600的方塊圖，其可為分立的圖形處理單元、也可以是與多個處理核心或其他半導體裝置(諸如但不限於記憶體裝置或網路介面)整合的圖形處理器。圖形處理器1600可為圖形處理器1508的變化，

並且可以用來替代圖形處理器 1508。因此，本文結合圖形處理器 1508 的任何特徵之揭露亦揭露了與圖形處理器 1600 的相應組合，但不限於此。圖形處理器可經由記憶體映射的 I/O 介面與圖形處理器上之暫存器以及放置在處理器記憶體內的命令進行通訊。圖形處理器 1600 可包括用以存取記憶體的記憶體介面 1614。記憶體介面 1614 可以是對於本地記憶體、一或多個內部快取、一或多個共享外部快取及/或對於系統記憶體之介面。

【0267】可選地，圖形處理器 1600 還包括用以驅動顯示輸出資料至顯示裝置 1618 之顯示控制器 1602。顯示控制器 1602 包括用於多層視訊或使用者介面元件的顯示及組成的一或多個重疊平面之硬體。顯示裝置 1618 可為內部或外部顯示裝置。在一實施例中，顯示裝置 1618 係頭戴式顯示裝置，諸如虛擬實境 (VR) 顯示裝置或擴增實境 (AR) 顯示裝置。圖形處理器 1600 包括視訊編碼解碼引擎 1606，用以將媒體編碼、解碼或轉碼至、自或介於一或多個媒體編碼格式之間，包括但不限於動態圖像專家群組 (MPEG) 格式 (諸如 MPEG-2)、先進視訊編碼 (AVC) 格式 (諸如 H.264/MPEG-4 AVC、H.265/HEVC)、開放媒體聯盟 (Alliance for Open) (AOMedia) VP8、VP9、以及電影電視工程師協會 (SMPTE) 421M/VC-1 及聯合圖像專家群組 (JPEG) 格式 (諸如 JPEG 和動態 JPEG (MJPEG) 格式)。

【0268】圖形處理器 1600 可包括區塊影像轉移 (BLIT) 引擎 1603，用以執行二維 (2D) 光柵化器操作，包括 (例如)

位元邊界區塊轉移。然而，2D圖形操作係使用圖形處理引擎(GPE)1610中的一或多個組件來執行。在一些實施例中，GPE 1610為用以執行包括三維(3D)圖形操作及媒體操作之圖形操作的計算引擎。

【0269】 GPE 1610可包括用以執行諸如使用作用於3D基元形狀(例如，矩形、三角形，等等)上之處理功能以渲染三維影像及場景之3D操作的3D管線1612。3D管線1612包括可編程及固定功能元件，其執行該元件內的各種工作及/或生產執行緒至3D/媒體子系統1615。雖然3D管線1612可被用來執行媒體操作，但GPE 1610之實施例還包括媒體管線1616，其被明確地用來執行媒體操作，諸如視訊後製處理及影像強化。

【0270】 媒體管線1616可包括固定功能或可編程邏輯單元，用以執行一或多個特殊化媒體操作，諸如視訊解碼加速、視訊去交錯及視訊編碼加速，以取代或代替視訊編碼解碼引擎1606。媒體管線1616可額外地包括執行緒生產單元以生產執行緒，以在3D/媒體子系統1615上執行。所生產的執行緒係執行針對3D/媒體子系統1615中所包括的一或多個圖形執行單元上的媒體操作之計算。

【0271】 3D/媒體子系統1615可包括用以執行由3D管線1612及媒體管線1616所生產的執行緒之邏輯。所述管線係傳送執行緒執行請求至3D/媒體子系統1615，其包括執行緒調度邏輯，用以將各個請求仲裁並調度至可用的執行緒執行資源。執行資源包括圖形執行單元之陣列，用以處

理 3D 及媒體執行緒。所述 3D/媒體子系統 1615 可包括用於執行緒指令及資料中的一或多個內部快取。額外地，3D/媒體子系統 1615 還可包括共享記憶體，包括暫存器及可定址記憶體，用以共享執行緒之間的資料並儲存輸出資料。

【0272】圖 16B 繪示圖形處理器 1620，並其可作為圖形處理器 1600 之變化並且可替代圖形處理器 1600，反之亦然。因此，本文結合圖形處理器 1600 的任何特徵之揭露亦揭露了與圖形處理器 1620 的相應組合，但不限於此。根據本文敘述的實施例，圖形處理器 1620 具有平鋪架構。圖形處理器 1620 可包括圖形處理引擎叢集 1622，其具有圖形引擎磚片 1610A-1610D 內圖 16A 之圖形處理引擎 1610 的多個實例。每一圖形引擎磚片 1610A-1610D 可經由一組磚片互連 1623A-1623F 互連。每一圖形引擎磚片 1610A-1610D 亦可經由記憶體互連 1625A-1625D 連接到記憶體模組或記憶體裝置 1626A-1626D。記憶體裝置 1626A-1626D 可使用任何圖形記憶體技術。例如，記憶體裝置 1626A-1626D 可為圖形雙資料率 (GDDR) 記憶體。記憶體裝置 1626A-1626D 可為高帶寬記憶體 (HBM) 模組，其可為與它們各別圖形引擎磚片 1610A-1610D 一起位於晶粒上。記憶體裝置 1626A-1626D 可以是堆疊的記憶體裝置，其可以堆疊在它們各別圖形引擎磚片 1610A-1610D 的頂部。每一圖形引擎磚片 1610A-1610D 與相關聯記憶體 1626A-1626D 可駐留在單獨的小晶片上，其接合至基底晶粒或基底基材上，如圖 24B-24D 進一步詳細敘述。

【0273】圖形處理器 1620 可與非均勻記憶體存取 (NUMA) 系統組態，其中記憶體裝置 1626A-1626D 與相關聯圖形引擎磚片 1610A-1610D 耦接。給定的記憶體裝置可能會被圖形引擎磚片存取，而不是直接與其連接的磚片。然而，當存取本地磚片時，對存取記憶體裝置 1626A-1626D 的存取等待時間可以是最低的。在一實施例中，啟用快取一致性 NUMA(ccNUMA) 系統，該系統使用磚片互連 1623A-1623F 來啟用圖形引擎磚片 1610A-1610D 中的快取控制器之間的通訊，以在當超過一個快取儲存相同的記憶體位置時保持一致的記憶體影像。

【0274】圖形處理引擎叢集 1622 可與晶載或封裝上結構互連 1624 連接。在一實施例中，結構互連 1624 包括網路處理器、晶片上網路 (NoC)、或另一切換處理器，以使結構互連 1624 作為在圖形處理器 1620 的組件之間切換資料封裝的封裝切換式結構互連。結構互連 1624 可啟用圖形引擎磚片 1610A-1610D 與諸如視訊編解碼引擎 1606 和一或多個複製引擎 1604 的組件之間的通訊。複製引擎 1604 可用來將資料移出、移入記憶體裝置 1626A-1626D 和圖形處理器 1620 外部的記憶體 (例如，系統記憶體)，以及在記憶體裝置 1626A-1626D 和圖形處理器 1620 之間移動。結構互連 1624 還可用於互連圖形引擎磚片 1610A-1610D。圖形處理器 1620 可選地包括顯示控制器 1602，其實現與外部顯示裝置 1618 的連接。圖形處理器亦可組態為圖形或計算加速器。在加速器的組態中，可省略顯示控制器 1602 和顯示裝

置 1618。

【0275】圖形處理器 1620 可經由主介面 1628 連接至主機系統。主介面 1628 可以啟用圖形處理器 1620、系統記憶體、及/或其他系統組件之間的通訊。主介面 1628 可為(例如)PCI express 匯流排或另一類型的主機系統介面。例如，主介面 1628 可以是 NVLink 或 NVSwitch 介面。主介面 1628 及結構互連 1624 可以協作以使圖形處理器 1620 的多個實例能夠作為單一邏輯裝置。主介面 1628 及結構互連 1624 之間的合作還可以使每一圖形引擎磚片 1610A-1610D 作為不同的邏輯圖形裝置被呈現給主機系統。

【0276】圖 16C 繪示根據本文敘述之實施例的計算加速器 1630。計算加速器 1630 可包括與圖 16B 的圖形處理器 1620 的架構相似並且被優化用於計算加速。計算引擎叢集 1632 可包括一組計算引擎磚片 1640A-1640D，其包括針對並行或基於向量的通用計算操作而優化的執行邏輯。儘管在一些實施例中，計算引擎磚片 1640A-1640D 中的一或多個可以包括執行媒體加速的邏輯，但是計算引擎磚片 1640A-1640D 可以不包括固定功能圖形處理邏輯。計算引擎磚片 1640A-1640D 可以經由記憶體互連 1625A-1625D 連接到記憶體 1626A-1626D。記憶體 1626A-1626D 和記憶體互連 1625A-1625D 可以與圖形處理器 1620 中的技術類似，也可以不同。圖形計算引擎磚片 1640A-1640D 也可以經由一組磚片互連 1623A-1623F 互連，並且可以與結構互連 1624 連接及/或藉由結構互連 1624 互連。在一實施例中，

計算加速器 1630 包括大的 L3 快取 1636，其可經組態為裝置寬度快取。計算加速器 1630 還可經由以如圖 16B 的圖形處理器 1620 類似方式的主介面 1628 連接到主機處理器和記憶體。

【0277】計算加速器 1630 還可以包括積體網路介面 1642。在一實施例中，網路介面 1642 包括網路處理器和控制器邏輯，其使得計算引擎叢集 1632 能夠在實體層互連 1644 上進行通訊而無需資料遍歷主機系統的記憶體。在一實施例中，計算引擎磚片 1640A-1640D 中的一者被網路處理器邏輯所取代，並經由實體層互連 1644 傳輸或接收的資料可以直接地傳輸至記憶體 1626A-1626D 或從記憶體 1626A-1626D 傳輸。可以經由實體層互連 1644 將計算加速器 1630 的多個實例結合到單一邏輯裝置中。替代地，各種計算引擎磚片 1640A-1640D 可以呈現為不同的網路可存取的計算加速器裝置。

圖形處理引擎

【0278】圖 17 係根據一些實施例之圖形處理器之圖形處理引擎 1710 的方塊圖。圖形處理引擎 (GPE) 1710 可以是圖 16A 所示的 GPE 1610 的一種版本，並且還可以表示圖 16B 的圖形引擎磚片 1610A-1610D。具有與本文中任何其他圖式之元件相同或相似名稱之圖 17 的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實

體，但是不限於此。例如，圖 16A 之 3D 管線 1612 及媒體管線 1616 亦繪示在圖 17。媒體管線 1616 在 GPE 1710 的一些實施例中是選擇性的，且可能不被明確地包括在 GPE 1710 內。例如以及在至少一個實施例中，個別的媒體及/或影像處理器被耦接至 GPE 1710。

【0279】 GPE 1710 可與命令串流器 1703 耦接或包括命令串流器 1703，命令串流器 1703 提供命令串流至 3D 管線 1612 及/或媒體管線 1616。替代地或額外地，命令串流器 1703 可直接地耦接至統一返回緩衝器 1718。統一返回緩衝器 1718 可通訊地耦接至圖形核心陣列 1714。可選地，命令串流器 1703 係與記憶體耦接，該記憶體可以是系統記憶體或內部快取記憶體及共享快取記憶體中的一或多者。命令串流器 1703 可從記憶體接收命令並將該些命令傳送至 3D 管線 1612 及/或媒體管線 1616。該些命令被直接提取自環狀緩衝器，其儲存 3D 管線 1612 及媒體管線 1616 之命令。環狀緩衝器可額外地包括批次命令緩衝器，其儲存多個命令之批次。3D 管線 1612 之命令還可包括對於記憶體中所儲存之資料的參考，諸如但不限於用於 3D 管線 1612 之頂點和幾何資料及/或用於媒體管線 1616 之影像資料和記憶體物件。3D 管線 1612 及媒體管線 1616 係藉由經由個別管線內的邏輯以執行操作或將一或多個執行緒調度至圖形核心陣列 1714 來處理該些命令及資料。圖形核心陣列 1714 可以包括一或多個圖形核心區塊(例如，圖形核心 1715A、圖形核心 1715B)，每一區塊包括一或多個圖形核心。每一圖形核心

包括一組圖形執行資源，其包括用於執行圖形和計算操作的通用和圖形特定執行邏輯，以及固定功能紋理處理及/或機器學習以及人工智能加速邏輯。

【0280】在各種實施例中，3D管線1612可包括固定功能和可編程邏輯，其用以執行一或多個著色器程式，諸如頂點著色器、幾何著色器、像素著色器、片段著色器、計算著色器或其他著色器程式，藉由處理該些指令並將執行執行緒調度至圖形核心陣列1714。圖形核心陣列1714提供用於執行這些著色器程式的統一執行資源區塊。在圖形核心陣列1714之圖形核心1715A-1715B內的多用途執行邏輯(例如，執行單元)包括支持各種3D API著色器語言並可執行與多個著色器相關的多個同時執行執行緒。

【0281】圖形核心陣列1714可包括用以執行諸如視訊及/或影像處理之媒體功能的執行邏輯。執行單元可額外地包括通用邏輯，其可編程以執行除了圖形處理操作之外的並行通用計算操作。通用邏輯可並行地或聯合圖14之處理器核心1407或圖15A中之核心1502A-1502N內的通用邏輯來執行處理操作。

【0282】由圖形核心陣列1714上所執行之執行緒所產生的輸出資料可將資料輸出至統一返回緩衝器(URB)1718中之記憶體。URB 1718可儲存多個執行緒之資料。URB 1718可被用來傳送資料於圖形核心陣列1714上所執行的執行緒之間。URB 1718可額外地被用於圖形核心陣列1714上的執行緒與共享功能邏輯1720內的固定功能邏輯之間的

同步化。

【0283】可選地，圖形核心陣列 1714 可為可縮放的，以致使該陣列包括可變數目的圖形核心，其根據 GPE 1710 之目標功率及效能位準而各具有可變數目的執行單元。執行資源可為動態可縮放的，以致使執行資源可根據所需來啟用或禁用。

【0284】圖形核心陣列 1714 係與共享功能邏輯 1720 耦接，共享功能邏輯 1720 包括被共享於圖形核心陣列中的圖形核心之間的多個資源。在共享功能邏輯 1720 之中的共享功能為硬體邏輯單元，其提供特殊化補充功能給圖形核心陣列 1714。在各種實施例中，共享功能邏輯 1720 包括但不限於取樣器邏輯 1721、數學邏輯 1722 及執行緒間通訊 (ITC) 邏輯 1723。額外地，可實施共享功能邏輯 1720 內的一或多個快取 1725。

【0285】共享功能至少被實現在針對既定特殊化功能之需求的情況下，即使包括在圖形核心陣列 1714 內仍不足時。取而代之地，該特殊化功能之單一例示被實現為共享功能邏輯 1720 中之獨立單體且被共享於圖形核心陣列 1714 內的執行資源之間。被共享於圖形核心陣列 1714 之間且被包括在圖形核心陣列 1714 內的功能的精確集合係在實施例之間改變。圖形核心陣列 1714 廣泛使用的共享功能邏輯 1720 內的特定共享功能可以包括在圖形核心陣列 1714 內的共享功能邏輯 1716 內。可選地，圖形核心陣列 1714 內的共享功能邏輯 1716 可以包括共享功能邏輯 1720 內的一些或全

部的邏輯。共享功能邏輯 1720 內的所有邏輯元件可以在圖形核心陣列 1714 的共享功能邏輯 1716 內複製。替代地，共享功能邏輯 1720 被排除在外，從而有利於圖形核心陣列 1714 內的共享功能邏輯 1716。

執行單元

【0286】圖 18A-18B 繪示根據本文敘述的實施例之包括在圖形處理器核心中採用之處理元件陣列的執行緒執行邏輯 1800。具有與本文中任何其他圖式之元件相同或相似名稱之圖 18A-18B 的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實體，但是不限於此。圖 18A-18B 繪示執行緒執行邏輯 1800 的概述，其可代表圖 15B 的每一子核心 1521A-1521F 所示的硬體邏輯。圖 18A 係通用圖形處理器內之執行單元的代表圖，而圖 18B 係可在計算加速器內使用之執行單元的代表圖。

【0287】如圖 18A 所繪示，執行緒執行邏輯 1800 可包括著色器處理器 1802、執行緒分派器 1804、指令快取 1806、包括複數個圖形執行單元 1808A-1808N 的可縮放執行單元陣列、取樣器 1810、共享本地記憶體 1811、資料快取 1812、及資料埠 1814。可選地，可縮放執行單元陣列可基於工作負載的計算要求啟用或禁用一或多個執行單元(例如，執行單元 1808A、1808B、1808C、1808D 至 1808N-1 及 1808N 中的任一者)來動態地縮放。所包括的組件可為

經由鏈接至該些組件之各者的互連結構來互連。執行緒執行邏輯 1800 可包括透過指令快取 1806、資料埠 1814、取樣器 1810、及圖形執行單元 1808A-1808N 中的一或多者至諸如系統記憶體或快取記憶體之記憶體的一或多個連接。各執行單元(例如，1808A)可為獨立可編程通用計算單元，其能夠執行多個同步硬體執行緒，而同時針對各執行緒並行地處理多個資料元件。在各種實施例中，執行單元 1808A-1808N 之陣列為可縮放的，以包括任何數目的個別執行單元。

【0288】在一些實施例中，圖形執行單元 1808A-1808N 可主要被用來執行著色器程式。著色器處理器 1802 可處理各種著色器程式，並經由執行緒調度器 1804 來調度與該些著色器程式相關的執行緒。執行緒分派器包括用以仲裁來自圖形和媒體管線之執行緒起始請求，並將該些請求的執行緒實例化於圖形執行單元 1808A-1808N 中的一或多個執行單元上之邏輯。例如，幾何管線可將頂點、鑲嵌或幾何著色器調度至執行緒執行邏輯以供處理。可選地，執行緒調度器 1804 還可處理來自執行中著色器程式的運行時間執行緒生產請求。

【0289】在一些實施例中，圖形執行單元 1808A-1808N 支持包括對於許多標準 3D 圖形著色器指令之本機支持的指令集，以致使來自圖形庫(例如，Direct 3D 及 OpenGL)之著色器程式以最少轉換被執行。執行單元支持頂點和幾何處理(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)及通用處理(例如,計算和媒體著色器)。圖形執行單元 1808A-1808N 之各者能夠多重發送單指令多資料(SIMD)執行,而多執行緒操作係致使在面對較高延遲時間記憶體存取時之有效率的執行環境。每一個執行單元內的各硬體執行緒具有專用的高頻寬暫存器檔案及相關的獨立執行緒狀態。執行係每個時脈多重發送至管線,其得以進行整數、單和雙精確度浮點操作、SIMD 分支能力、邏輯操作、超越操作及其他各種操作。當等待來自記憶體之資料或共享功能之一時,執行單元 1808A-1808N 內的相依性邏輯係致使等待執行緒休眠,直到該請求的資料已被返回。當該等待執行緒正在休眠時,硬體資源可被用於處理其他執行緒。例如,在與頂點著色器操作相關的延遲期間,執行單元可執行用於像素著色器、片段著色器或其他類型的著色器程式,包括不同的頂點著色器,諸如圖 21 所示的頂點著色器 2107。各種實施例可適用於藉由單指令多執行緒(SIMT)的使用來執行,以替代 SIMD 的使用或除了使用 SIMD 之外。參照 SIMD 核心或操作可以同時應用於 SIMT 或與 SIMT 一起應用於 SIMD。

【0290】圖形執行單元 1808A-1808N 中之各執行單元操作於資料元件之陣列上。資料元件的數目為「執行大小」或針對該指令之通道數目。執行通道為針對指令內的資料元件存取、遮蔽及流程控制的執行之邏輯單元。通道的數量可以獨立於特定圖形處理器的實體算術邏輯單元

(ALU)、浮點單元(FPU)或其他邏輯單元(例如，張量核心，射線追蹤核心等)的數量。額外地，圖形執行單元 1808A-1808N可支持整數和浮點資料類型。

【0291】執行單元指令集包括SIMD指令。各個資料元件可被儲存為暫存器中之緊縮資料類型，且執行單元將根據該些元件之資料大小來處理各個元件。例如，當操作於256位元寬的向量時，該向量的256位元被儲存於暫存器中，且執行單元係操作於該向量上而成為四個個別的184位元緊縮資料元件(四字組(QW)大小資料元件)、八個個別的32位元緊縮資料元件(雙字組(DW)大小資料元件)、十六個個別的16位元緊縮資料元件(字組(W)大小資料元件)或三十二個個別的8位元緊縮資料元件(位元組(B)大小資料元件)。然而，不同的向量寬度及暫存器大小是可能的。

【0292】可選地，可以將一或多個執行單元組合到具有對融合EU通用的執行緒控制邏輯(1807A-1807N)的融合圖形執行單元 1809A-1809N。可以將多個EU融合為一EU群組。可將融合EU群組中每一EU組態以執行單獨的SIMD硬體執行緒。在融合EU群組中EU的數量可根據實施例而變化。額外地，每個EU可以執行各種SIMD寬度，包括但不限於SIMD8、SIMD16、及SIMD32。每一融合圖形執行單元 1809A-1809N包括至少兩個執行單元。例如，融合執行單元 1809A包括第一EU 1808A、第二EU 1808B及第一EU 1808A和第二EU 1808B共有的執行緒控制邏輯 1807A。執行緒控制邏輯 1807A控制在融合圖形執行單元 1809A上

執行的執行緒，從而允許融合執行單元 1809A-1809N內的每一EU使用共享指令指示器暫存器來執行。

【0293】一或多個內部指令快取(例如，1806)被包括在執行緒執行邏輯1800中，以將執行單元之執行緒指令快取。一或多個資料快取(例如，1812)可被包括在執行緒執行邏輯1800中，以在執行緒執行期間快取執行緒資料。在執行邏輯1800上執行的執行緒亦可將顯示管理的資料儲存在共享本地記憶體1811。取樣器1810可被包括以提供針對3D操作之紋理取樣及針對媒體操作之媒體取樣。取樣器1810包括特殊化紋理或媒體取樣功能，用以在提供已取樣資料至執行單元之前，處理在取樣程序期間之紋理或媒體資料。

【0294】在執行期間，圖形及媒體管線係經由執行緒生產和調度邏輯來將執行緒起始請求傳送至執行緒執行邏輯1800。一旦幾何物件之群組已被處理並光柵化為像素資料，則著色器處理器1802內的像素處理器邏輯(例如，像素著色器邏輯、片段著色器邏輯，等等)被調用以進一步計算輸出資訊並致使結果被寫入至輸出表面(例如，顏色緩衝器、深度緩衝器、模板緩衝器，等等)。像素著色器或片段著色器可計算將跨該光柵化物件被內插的各個頂點屬性之值。著色器處理器1802內的像素處理器邏輯接著執行應用程式編程介面(API)供應的像素或片段著色器程式。為了執行著色器程式，著色器處理器1802經由執行緒調度器1804來將執行緒調度至執行單元(例如，1808A)。

著色器處理器 1802 可使用取樣器 1810 中之紋理取樣邏輯來存取記憶體中所儲存之紋理映射中的紋理資料。紋理資料及輸入幾何資料上的算術操作係計算各幾何片段之像素顏色資料或丟棄一或多個像素而不做進一步處理。

【0295】此外，資料埠 1814 可提供用於執行緒執行邏輯 1800 之記憶體存取機制，以將經處理資料輸出至記憶體以在圖形處理器輸出管線上進一步處理。資料埠 1814 可包括或耦接至一或多個快取記憶體(例如，資料快取 1812)以快取資料以經由資料埠 1814 進行記憶體存取。

【0296】可選地，執行邏輯 1800 亦可包括射線追蹤器 1805，其可提供射線追蹤加速功能。射線追蹤器 1805 可支持射線追蹤指令集，其包括用於射線產生的指令/功能。射線追蹤指令集可與圖 3 中的射線追蹤核心 372 所支持的射線追蹤指令集相似或不同。

【0297】圖 18B 繪示執行單元 1808 的例示性內部細節。圖形執行單元 1808 可包括指令提取單元 1837、通用暫存器檔案陣列 (GRF) 1824、架構暫存器檔案陣列 (ARF) 1826、執行緒仲裁器 1822、發送單元 1830、分支單元 1832、一組 SIMD 浮點單元 (FPU) 1834、以及可選地一組專用整數 SIMD ALU 1835。GRF 1824 和 ARF 1826 包括與每一同時在圖形執行單元 1808 內啟用之硬體執行緒相關聯的該組通用暫存器檔和架構暫存器檔。可以在 ARF 1826 中維持每一執行架構狀態，而在執行緒執行期間使用的資料係儲存在 GRF 1824 中。每一執行緒的執行狀態(包括每一

執行緒的指令指示器)可以保持在 ARF 1826中的執行緒指明暫存器。

【0298】圖形執行單元 1808可具有為同步多執行緒(SMT)及精細交錯多執行緒(IMT)之組合的架構。所述架構可具有模組化組態，其可在設計時間基於同時執行緒的目標數量和每個執行單元的暫存器數量進行微調，其中執行單元資源在用以執行多個同時執行緒的邏輯之間分配。圖形執行單元 1808可執行的邏輯執行緒的數量不限於硬體執行緒的數量，並且可將多個邏輯執行緒分配給每一硬體執行緒。

【0299】可選地，圖形執行單元 1808可以共同發布多個指令，每一指令可以是不同的指令。圖形執行單元執行緒 1808的執行緒仲裁器 1822可將指令分派給發送單元 1830、分支單元 1832、或 SIMD FPU 1834中的一者，以用於執行。每一執行緒可存取 GRF 1824內的 128通用暫存器，其中每一暫存器可儲存 32個位元組，可存取為 32位元資料元件的 SIMD 8元件向量。每一執行單元執行緒可存取 GRF 1824內的 4 Kbytes，儘管實施例不限於此，並且在其他實施例中可以提供更多或更少的暫存器資源。圖形執行單元 1808可劃分為可獨立地執行計算操作的七個硬體執行緒，儘管每一執行單元的執行緒數量也可以根據實施例而變化，例如，可以支持多達 16個硬體執行緒。在例示性實施例中，其中七個執行緒可存取 4 Kbytes，GRF 1824可總共儲存 28 Kbytes。在另一例示性實施例中，其中 16個執

行緒可存取 4 Kbytes，GRF 1824可總共儲存 64 Kbytes。然而，每個執行單元的執行緒數不限於那些實例，並且可以大於或小於給定的數目。靈活的定址模式可允許對暫存器進行一起定址，以有效地構建更寬的暫存器或表示跨步的矩形塊資料結構。

【0300】額外地或替代地，可經由由訊息傳遞發送單元 1830執行的「發送」指令來分派記憶體操作、取樣器操作、及其他更長等待時間的系統通訊。分支指令可分派至專用分支單元 1832，以促進 SIMD發散和最終收斂。

【0301】圖形執行單元 1808可包括一或多個 SIMD浮點單元(FPU)1834，以執行浮點操作。FPU 1834亦可支持整數計算。在一些情況下，FPU 1834可 SIMD執行多達 M個 32位元浮點(或整數)操作，或者 SIMD執行多達 2M個 16位元整數或 16位元浮點操作。可選地，至少一個 FPU提供擴展的數學功能以支持高通量超越數學函數及雙精度 184位元浮點。亦可存在一組 8位元整數 SIMD ALU 1835，並且可以對其進行特別優化，以執行與機器學習計算相關的操作。

【0302】可選地，圖形執行單元 1808之多個實例陣列可在圖形子核心分組(例如，子片)中實例化。為了實現可縮放性，產品架構者可以為每一子核心分組選擇確切的執行單元數量。執行單元 1808可橫跨複數個執行通道執行指令。此外，可在不同通道上執行在圖形執行單元 1808上執行的每一執行緒。

【0303】圖 19 繪示的進一步例示性執行單元 1900。具有與本文中任何其他圖式之元件相同或相似名稱之圖 19 的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實體，但是不限於此。執行單元 1900 可為用於例如圖 16C 中的計算引擎磚片 1640A-1640D 中計算優化的執行單元，但不限於此。如在圖 16B，執行單元 1900 亦可用在圖形引擎磚片 1610A-1610D 中。執行單元 1900 可包括執行緒控制單元 1901、執行緒狀態單元 1902、指令提取/預提取單元 1903、及指令解碼單元 1904。執行單元 1900 可額外地包括暫存器檔案 1906，其儲存可分配至執行單元內的硬體執行緒的暫存器。執行單元 1900 可額外地包括發送單元 1907 和分支單元 1908。發送單元 1907 和分支單元 1908 可以與圖 18B 之圖形執行單元 1808 的發送單元 1830 和分支單元 1832 類似地操作。

【0304】執行單元 1900 亦可包括計算單元 1910，其包括多個不同類型的功能單元。計算單元 1910 亦可包括 ALU 1911、脈動陣列 1912 和數學單元 1913。ALU 1911 包括算術邏輯單元的陣列。ALU 1911 可經組態以跨多個處理通道和資料通道以針對多個硬體及/或軟體執行緒執行 64 位元、32 位元、16 位元整數和浮點操作。ALU 1911 可同時執行整數和浮點操作(例如，在相同時脈循環)。

【0305】脈動陣列 1912 包括資料處理單元的 W 寬和 D 深的網路，其可用於以脈動方式執行向量和其他資料並行

操作。脈動陣列 1912 可經組態以執行各種矩陣操作，包括作為內積、外積、及通用矩陣-矩陣乘法(GEMM)操作。脈動陣列 1912 可支持 16 位元浮點操作，以及 8 位元、4 位元、2 位元和二進制整數操作。脈動陣列 1912 可組態以加速機器學習操作。脈動陣列 1912 可經組態為支持 bfloat16(大腦浮點數)16 位元浮點格式或張量浮點 32 位元浮點格式(TF32)，其尾數和指數位的數量相對於電機電子工程師學會(IEEE)754 格式不同。也支持 FP64 格式。

【0306】在一實施例中，脈動陣列 1912 包括用以加速稀疏矩陣操作的硬體。可以旁路輸入資料稀疏區域的乘法操作，而不會犧牲吞吐量。可以檢測輸入矩陣內的區塊稀疏性，並且可以繞開具有已知輸出值的操作。在一實施例中，脈動陣列 1912 包括使得能夠對具有壓縮表示的稀疏資料進行操作的硬體。稀疏矩陣的壓縮表示儲存非零值和元資料，其界定非零值在矩陣內的位置。例示性壓縮表示包括但不限於壓縮張量表示，例如壓縮稀疏列(CSR)、壓縮稀疏行(CSC)、壓縮稀疏纖維(CSF)表示。對壓縮表示的支持使操作能夠以壓縮張量格式在輸入上執行，而無需對壓縮表示進行解壓縮或解碼。在這樣的實施例中，可以僅對非零輸入值執行操作，並且可以將所得的非零輸出值映射到輸出矩陣。在一些實施例中，還提供針對機器專用無損資料壓縮格式的硬體支持，其在當硬體或跨系統匯流排內傳輸資料時使用。此資料可以以壓縮格式保留，以用於稀疏輸入資料，並且脈動陣列 1912 可以將壓縮元資料用於壓

縮資料，以使得能夠僅對非零值執行操作，或者使得能夠旁路零資料輸入的區塊而用於乘法操作。

【0307】與 ALU 單元 1911 相比，數學單元 1913 可組態為以高效且低功率的方式執行數學操作的特定子集。數學單元 1913 可以包括在所敘述的其他實施例提供的圖形處理引擎的共享函數邏輯中找到的數學邏輯，例如，圖 17 的共享函數邏輯 1720 的數學邏輯 1722。數學單元 1913 可被組態以執行 32 位元和 64 位元浮點操作。

【0308】執行緒控制單元 1901 包括用以控制執行單元內執行緒之執行的邏輯。執行緒控制單元 1901 可包括執行緒仲裁邏輯，其用以開始、停止、及先佔執行單元 1900 內的執行緒之執行。執行緒狀態單元 1902 可用以儲存分配給在執行單元 1900 上執行的執行緒的執行緒狀態。當執行緒狀態為阻塞或閒置時，將執行緒狀態儲存在執行緒單元 1900 中使得能夠快速先佔執行緒。指令提取/預提取單元 1903 可以從更高級執行邏輯的指令快取(例如，如圖 18A 中的指令快取 1806)提取指令。指令提取/預提取單元 1903 亦可基於對當前執行的執行緒分析，發出對要加載至指令快取中之指令的預提取請求。指令解碼單元 1904 可用來解碼將由計算單元執行的指令。指令解碼單元 1904 可以用作輔助解碼器，以將複雜的指令解碼為組成的微操作。

【0309】執行單元 1900 額外地包括暫存器檔案 1906，其可由在執行單元 1900 上執行的硬體執行緒使用。暫存器檔案 1906 中的暫存器可以在用於執行執行單元 1900 之計算

單元 1910 內的多個同時執行緒的邏輯上劃分。圖形執行單元 1900 可執行的邏輯執行緒的數量不限於硬體執行緒的數量，並且可將多個邏輯執行緒分配給每一硬體執行緒。暫存器檔案 1906 的大小可以基於支持的硬體執行緒的數量而在個實施例中變化。暫存器重新命名可用來將暫存器動態地分配給硬體執行緒。

【0310】圖 20 係繪示圖形處理器指令格式 2000 的方塊圖。圖形處理器執行單元係支持一種具有多個格式之指令的指令集。實線方盒係顯示通常地被包括在執行單元指令中之組件，而虛線則包括為選擇性的或者僅被包括在該些指令之子集中的組件。在一些實施例中，將圖形處理器指令格式 2000 描述並說明為巨集指令，一旦該指令被處理後，其為供應至執行單元之指令，而不是得自指令解碼之微操作。因此，單一指令可導致硬體執行多個微操作。

【0311】如本文所述之圖形處理器執行單元可本地地支持 128 位元指令格式 2010 的指令。64 位元壓縮指令格式 2030 可根據選定的指令、指令選項及操作元的數目而用於一些指令。本質 128 位元指令格式 2010 係提供對於所有指令選項之存取，而一些選項及操作被侷限於 64 位元格式 2030。可用於 64 位元格式 2030 之本質指令隨實施例而改變。所述指令係使用指標欄位 2013 中之一組指標值而被部分地壓縮。執行單元硬體係根據指標值而參考一組壓縮表，並使用壓縮表輸出以重新建構 128 位元指令格式 2010 之本質指令。可使用指令的其他大小和格式。

【0312】針對各格式，指令運算碼2012係界定該執行單元應執行之操作。執行單元跨各操作元之多資料元件來並行地執行各指令。例如，回應於加法指令，執行單元跨代表紋理元件或圖片元件之各顏色通道來執行同步加法操作。預設地，執行單元跨操作元的所有資料通道來執行各指令。指令控制欄位2014可致使對於某些執行選項之控制，諸如通道選擇(例如，斷定)及資料通道順序(例如，拌合)。針對128位元指令格式2010之指令，執行大小欄位2016係限制將被並行地執行之資料通道的數目。執行大小欄位2016可能無法在64位元緊湊指令格式2030中使用。

【0313】一些執行單元指令具有高達三操作元，其包括兩個來源操作元(src0 2020、src1 2022)及一個目的地2018。執行單元可支持雙目的地指令，其中該些目的地之一者被隱含。資料操縱指令可具有第三來源操作元(例如，SRC2 2024)，其中指令運算碼2012係判定來源操作元的數目。指令的最後來源操作元可以是利用該指令傳遞的立即(例如，硬編碼)值。

【0314】128位元指令格式2010可包括存取/位址模式欄位2026，其指明(例如)直接暫存器定址模式或是間接暫存器定址模式被使用。當直接暫存器定址模式被使用時，一或多個操作元之暫存器位址係直接地由該指令中之位元來提供。

【0315】128位元指令格式2010還可包括存取/位址模式欄位2026，其指明指令的位址模式及/或存取模式。存

取模式可被用來界定該指令之資料存取對準。可支持包括 16 位元組對準的存取模式以及 1 位元組對準的存取模式，其中存取模式之位元組對準係判定指令操作元之存取對準。例如，當在第一模式時，該指令對於來源和目的地操作元可使用位元組對準的定址，而當在第二模式時，該指令對於來源和目的地操作元可使用 16 位元組對準的定址。

【0316】存取/位址模式欄位 2026 之位址模式部分可判定該指令是否使用直接或者間接定址。當直接暫存器定址模式被使用時，該指令中之位元係直接地提供一或多個操作元之暫存器位址。當間接暫存器定址模式被使用時，一或多個操作元之暫存器位址可根據該指令中之位址暫存器值及位址立即欄位而被計算。

【0317】指令係可基於運算碼 2012 位元欄位而被群組以簡化運算碼解碼 2040。針對 8 位元運算碼，位元 4、5 及 6 容許執行單元判定運算碼之類型。所示之精確運算碼群組僅為範例。移動和邏輯運算碼群組 2042 可包括資料移動和邏輯指令(例如，移動(mov)、比較(cmp))。移動和邏輯群組 2042 可共享五個最低有效位元(LSB)，其中移動(mov)指令為 0000xxxxb 之形式，而邏輯指令為 0001xxxxb 之形式。流程控制指令群組 2044(例如，呼叫、跳躍(jmp))包括為 0010xxxxb(例如，0x20)之形式的指令。雜項指令群組 2046 包括指令之混合，其包括為 0011xxxxb(例如，0x30)之形式的同步化指令(例如，等待、傳送)。並行數學指令群組 2048 包括為 0100xxxxb(例如，0x40)之形式的組件式

算術指令(例如，加法、乘法(mul))。並行數學指令群組 2048 係跨資料通道而並行地執行算術操作。向量數學群組 2050 包括為 $0101xxxxb$ (例如， $0x50$)之形式的算術指令(例如，dp4)。向量數學群組係執行諸如對於向量操作元之內積計算之算術。在一實施例中，所繪示之運算碼解碼 2040 可用來判定執行單元中的哪個部分將用於執行解碼的指令。例如，一些指令可被指定為將由脈動陣列執行的脈動指令。其他指令(諸如，射線追蹤指令(未圖示))可以路由至執行邏輯之切片或分區內的射線追蹤核心或射線追蹤邏輯。

圖形管線

【0318】圖 21 係根據另一實施例之圖形處理器 2100 的方塊圖。具有與本文中任何其他圖式之元件相同或相似名稱之圖 21 的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實體，但是不限於此。

【0319】圖形處理器 2100 可包括不同類型的圖形處理管線，諸如幾何管線 2120、媒體管線 2130、顯示引擎 2140、執行緒執行邏輯 2150、及渲染輸出管線 2170。圖形處理器 2100 可為包括一或多個通用處理核心之多核心處理系統內的圖形處理器。圖形處理器可為由暫存器寫入至一或多個控制暫存器(未圖示)所控制，或者經由發送至圖形處理器 2100(經由環狀互連 2102)之命令來控制。環狀互連

2102可將圖形處理器2100耦接到其他處理組件，諸如其他圖形處理器或通用處理器。來自環形互連2102的命令由命令串流器2103解譯，該命令串流器2103將指令提供給幾何管線2120或媒體管線2130的各個組件。

【0320】命令串流器2103可以指引頂點提取器2105的操作，該頂點提取器2105從記憶體讀取頂點資料並執行命令串流器2103提供的頂點處理命令。頂點提取器2105可提供頂點資料至頂點著色器2107，其對於每個頂點執行座標空間轉換及照亮操作。頂點提取器2105及頂點著色器2107可以藉由經由執行緒分配器2131將執行緒分配給執行單元2152A-2152B來執行頂點處理指令。

【0321】執行單元2152A-2152B可為具有用以執行圖形及媒體操作之指令集的向量處理器之陣列。執行單元2152A-2152B可以具有專用於每個陣列或在陣列之間共享的附接的L1快取2151。快取可被組態成資料快取、指令快取或單一快取，其被分區成在不同的分區中含有資料及指令。

【0322】幾何管線2120可包括鑲嵌組件，用以執行3D物件之硬體加速鑲嵌。可編程外殼著色器2111可組態鑲嵌操作。可編程領域著色器2117可提供鑲嵌輸出的後端評估。鑲嵌器2113可在外殼著色器2111的方向上操作，並含有專用的邏輯以基於作為幾何管線2120的輸入提供的粗略幾何模型產生一組詳細的幾何物件。另外，如果不使用鑲嵌，則可旁路鑲嵌組件(例如，外殼著色器2111、鑲嵌器

2113、及領域著色器 2117)。鑲嵌組件可基於從頂點著色器 2107 接收的資料來操作。

【0323】完整的幾何物件可由幾何著色器 2119 經由分配給執行單元 2152A-2152B 的一或多個執行緒來處理，或者可直接進入截波器 2129。幾何著色器可操作於整個幾何物件上，而非如圖形管線中的先前階段中之頂點或頂點的補丁。假如鑲嵌被禁用，則幾何著色器 2119 接收來自頂點著色器 2107 之輸入。假如鑲嵌單元被禁用，則幾何著色器 2119 可由幾何著色器程式所編程，以執行幾何鑲嵌。

【0324】在光柵化之前，截波器 2129 處理頂點資料。截波器 2129 可為固定功能截波器或具有截波及幾何著色器功能之可編程截波器。渲染輸出管線 2170 中的光柵化器和深度測試組件 2173 可分派像素著色器，以將幾何物件轉換成每一像素表示。像素著色器邏輯可包括在執行緒執行邏輯 2150 中。可選地，應用程式可省略光柵化器及深度測試組件 2173，並經由串流輸出單元 2123 來存取未光柵化的頂點資料。

【0325】圖形處理器 2100 具有互連匯流排、互連結構或一些其他互連機制，其容許資料及訊息傳遞於處理器的主要組件之間。在一些實施例中，執行單元 2152A-2152B 及相關聯的邏輯單元(例如，L1 快取 2151、取樣器 2154、紋理快取 2158 等)經由資料埠 2156 互連以執行記憶體存取並與處理器的渲染輸出管線組件通訊。取樣器 2154、快取 2151、2158 及執行單元 2152A-2152B 可以分別具有單獨的

記憶體存取路徑。可選地，紋理快取 2158 亦可組態為取樣器快取。

【0326】渲染輸出管線 2170 可含有光柵化器及深度測試組件 2173，其將頂點為基的物件轉換為相關之像素為基的表示。光柵化器邏輯可包括視窗器/遮蔽器單元，用以執行固定功能三角及直線光柵化。相關的渲染快取 2178 及深度快取 2179 還可用在一些實施例中。像素操作組件 2177 係對於資料執行像素為基的操作，雖然在一些例子中，與 2D 操作相關的像素操作(例如，利用混合之位元區塊影像轉移)係由 2D 引擎 2141 來執行或者於顯示時刻使用重疊顯示平面由顯示控制器 2143 來取代。共享 L3 快取 2175 可為用於所有圖形組件，其容許資料之共享而不使用主系統記憶體。

【0327】媒體管線 2130 可包括媒體引擎 2137 及視訊前端 2134。視訊前端 2134 可從命令串流器 2103 接收管線命令。媒體管線 2130 可包括單獨的命令串流器。視訊前端 2134 可以在將命令發送到媒體引擎 2137 之前處理媒體命令。媒體引擎 2137 可以包括執行緒產生功能，以產生執行緒以經由執行緒分派器 2131 分派給執行緒執行邏輯 2150。

【0328】圖形處理器 2100 可以包括顯示引擎 2140。顯示引擎 2140 可以在處理器 2100 的外部，並且可以經由環形互連 2102 或某個其他互連匯流排或結構與圖形處理器耦接。顯示引擎 2140 可以包括 2D 引擎 2141 和顯示控制器 2143。顯示引擎 2140 可含有專用邏輯，其能夠獨立自 3D 管

線而操作。顯示控制器 2143 可與顯示裝置(未顯示)耦接，顯示裝置可以是系統整合顯示裝置(如在膝上型電腦中)或經由顯示裝置連接器而附接的外部顯示裝置。

【0329】幾何管線 2120 及媒體管線 2130 可組態成基於多個圖形及媒體編程介面來執行操作，而非專用於任一應用程式編程介面(API)。圖形處理器之驅動器軟體可將專用於特定圖形或媒體庫的 API 呼叫轉換為可由圖形處理器所處理的命令。提供給開放式圖形庫(OpenGL)、開放式計算語言(OpenCL)及/或 Vulkan 圖形和計算 API 之支持，其均來自 Khronos 集團。還可提供支持給來自微軟公司的 Direct3D 庫。可支持這些庫之組合。還可提供支持給開放式來源電腦視覺庫(OpenCV)。假如可從未來 API 之管線執行至圖形處理器之管線的映射，則具有可相容 3D 管線之未來 API 亦將被支持。

圖形管線編程

【0330】圖 22A 係繪示用於編程圖形處理管線(諸如(例如)本文結合圖 16A、17、21 所述的管線)之圖形處理器命令格式 2200 的方塊圖。圖 22B 係根據一實施例繪示之圖形處理器命令序列 2210 的方塊圖。圖 22A 中之實線方盒係繪示通常被包括在圖形命令中之組件，而虛線則包括為選擇性的或者其僅被包括在圖形命令之子集中的組件。圖 22A 之例示性圖形處理器命令格式 2200 包括用於識別客戶端 2202 的資料欄位、命令運算碼(opcode) 2204 及用於命令

的資料2206。子運算碼2205及命令大小2208亦被包括在一些命令中。

【0331】客戶端2202指明處理該命令資料之圖形裝置的客戶端單元。圖形處理器命令解析器可檢查各命令之客戶欄位以調適該命令的進一步處理並將命令資料發送至適當的客戶端單元。圖形處理器客戶端單元可包括記憶體介面單元、渲染單元、2D單元、3D單元及媒體單元。每一個客戶端單元具有處理該些命令之相應處理管線。一旦該命令由客戶端單元所接收，客戶端單元便讀取運算碼2204，並且(假如存在的話)子運算碼2205用以判定應執行的操作。客戶端單元係使用資料欄位2206中之資訊來執行該命令。對於一些命令，預期有明確的命令大小2208以指明命令之大小。命令解析器可自動地根據命令運算碼來判定至少一些命令的大小。命令可經由多個雙字組來對準。亦可使用其他命令格式。

【0332】圖22B中之流程圖繪示了例示性圖形處理器命令序列2210。特徵在於例示性圖形處理器的資料處理系統的軟體或韌體可使用了所顯示之命令序列的版本以設定、執行及終止一組圖形操作。樣本命令序列被顯示並描述以僅供範例之目的，並不限於這些特定命令或此命令序列。此外，該些命令可被發送為命令序列中之命令的批次，以致使圖形處理器將以至少部分並行性來處理命令之序列。

【0333】圖形處理器命令序列2210可開始於管線清除

命令 2212，用以致使任何活動圖形管線完成該管線的目前擱置命令。可選地，3D管線 2222及媒體管線 2224可能無法同時操作。管線清除被執行，以致使活動圖形管線完成任何擱置的命令。回應於管線清除，圖形處理器之命令解析器將暫停命令處理，直到活動繪圖引擎完成擱置的操作且相關讀取快取被無效化。選擇性地，渲染快取中被標記為「髒」的任何資料可被清除至記憶體。管線清除命令 2212可被使用於管線同步化，或者在將圖形處理器置入低功率狀態之前。

【0334】當命令序列需要圖形處理器來明確地在管線之間切換時，管線選擇命令 2213可被使用。在發送管線命令之前，管線選擇命令 2213可能在執行背景內僅需要一次，除非該背景將發送命令給兩管線。可能在經由管線選擇命令 2213進行管線切換之前緊接需要管線清除命令 2212。

【0335】管線控制命令 2214可組態用於操作的圖形管線，並且可用於對3D管線 2222和媒體管線 2224進行編程。管線控制命令 2214可組態活動管線之管線狀態。管線控制命令 2214被用於管線同步化，並用以在處理命令之批次之前清除來自活動管線內的一或多個快取記憶體的資料。

【0336】與返回緩衝器狀態 2216有關的命令可用於為一組返回緩衝器組態個別管線來寫入資料。一些管線操作需要一或多個返回緩衝器之配置、選擇或組態，其中該些操作在處理期間將中間資料寫入該些返回緩衝器。圖形處

理器也使用一或多個返回緩衝器來儲存輸出資料並執行跨執行緒通訊。返回緩衝器狀態 2216 可包括選擇返回緩衝器的大小與數目以用於一組管線操作。

【0337】命令序列中的殘留命令係根據針對操作之活動管線而不同。基於管線判定 2220，命令序列被定制為以 3D 管線狀態 2230 開始的 3D 管線 2222 或以媒體管線狀態 2240 開始的媒體管線 2224。

【0338】組態 3D 管線狀態 2230 之命令包括用於頂點緩衝器狀態、頂點元件狀態、恆定顏色狀態、深度緩衝器狀態及其他狀態變數的 3D 狀態設定命令，其應在 3D 基元命令被處理之前被組態。這些命令之值係至少部分地根據使用中之特定 3D API 來判定。3D 管線狀態 2230 命令亦能夠選擇性地禁用或省略一些管線元件，假如那些元件將不被使用的話。

【0339】3D 基元 2232 命令可用於傳遞要由 3D 管線處理的 3D 基元。經由 3D 基元 2232 命令而被傳遞至圖形處理器之命令及相關參數被遞送至圖形管線中之頂點提取功能。頂點提取功能係使用 3D 基元 2232 命令資料來產生頂點資料結構。頂點資料結構被儲存於一或多個返回緩衝器中。3D 基元 2232 命令被用來經由頂點著色器來對於 3D 基元執行頂點操作。為了處理頂點著色器，3D 管線 2222 將著色器執行緒調度至圖形處理器執行單元。

【0340】3D 管線 2222 可經由執行 2234 命令或事件觸發。暫存器可寫入觸發命令執行。執行係經由命令序列中

之「去(go)」或「踢(kick)」命令而被觸發。命令執行係使用管線同步化命令而被觸發以透過圖形管線來清除該命令序列。3D管線將執行針對3D基元之幾何處理。一旦操作完成，所得的幾何物件被光柵化且像素引擎將所得像素上色。用以控制像素著色及像素後端操作之額外命令還可被包括以用於那些操作。

【0341】當執行媒體操作時，圖形處理器命令序列2210可以遵循媒體管線2224的路徑。通常，針對媒體管線2224之編程的特定使用及方式係取決於待執行之媒體或計算操作。在媒體解碼期間，特定媒體解碼操作可被卸載至媒體管線。媒體管線也可被省略，而媒體解碼可使用由一或多個通用處理核心所提供的資源而被整體地或部分地執行。媒體管線還包括用於通用圖形處理器單元(GPGPU)操作之元件，其中圖形處理器通常使用並非明確地相關於圖形基元之渲染的計算著色器程式來執行SIMD向量操作。

【0342】媒體管線2224可以以與3D管線2222類似的方式組態。在媒體物件命令2242之前，用以組態媒體管線狀態2240的一組命令被調度或置入命令佇列。媒體管線狀態2240的命令包括用以組態將被用來處理媒體物件的媒體管線元件之資料。此包括用以組態媒體管線內的視訊解碼及視訊編碼邏輯的資料，諸如編碼或解碼格式。媒體管線狀態2240的命令亦支持使用一或多個指向含有狀態設定之批次的「間接」狀態元件之指標。

【0343】媒體物件命令2242可供應指向媒體物件之指

標以供媒體管線處理。媒體物件包括含有待處理之視訊資料的記憶體緩衝器。可選地，在發送媒體物件命令 2242 之前，所有媒體管線狀態需為有效的。一旦管線狀態被組態且媒體物件命令 2242 被排列，則媒體管線 2224 係經由執行命令 2244 或同等執行事件(例如，暫存器寫入)而被觸發。來自媒體管線 2224 之輸出可接著藉由 3D 管線 2222 或媒體管線 2224 所提供的操作來後處理。GPGPU 操作可以以如媒體操作之類似方式來組態與執行。

圖形軟體架構

【0344】圖 23 繪示用於資料處理系統 2300 之例示性圖形軟體架構。此種軟體架可包括 3D 圖形應用程式 2310、作業系統 2320、以及至少一處理器 2330。處理器 2330 可以包括圖形處理器 2332 以及一或多個通用處理器核心 2334。處理器 2330 可以是處理器 1402 或本文敘述的任何其他處理器的變化。可以使用處理器 2330 取代處理器 1402 或本文敘述的任何其他處理器。因此，與處理器 1402 或本文敘述的任何其他處理器組合之任何特徵的揭露亦揭露了與圖形處理器 2330 的相應組合，但不限於此。除此之外，具有與本文中任何其他圖式之元件相同或相似名稱之圖 23 的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實體，但是不限於此。圖形應用程式 2310 與作業系統 2320 分別在資料處理系統的系統記憶體 2350 中執行。

【0345】3D圖形應用程式2310可包括一或多個著色器程式，包括著色器指令2312。著色器語言指令可以是高階著色器語言，諸如Direct3D之高階著色器語言(HLSL)或OpenGL著色器語言(GLSL)等。應用程式還可包括以適於由通用處理器核心2334執行的機器語言的可執行指令2314。應用程式還可包括由頂點資料所界定的圖形物件2316。

【0346】作業系統2320是來自微軟公司的Microsoft® Windows®作業系統、專用UNIX類作業系統或使用Linux內核之變化的開放式來源UNIX類作業系統。作業系統2320可支持圖形API 2322，諸如Direct3D API、OpenGL API或Vulkan API。當使用Direct3D API時，作業系統2320係使用前端著色器編譯器2324來將HLSL中的任何著色器指令2312編譯為低階著色器語言。該編譯可以是及時(JIT)編譯，或者該應用程式可執行著色器預編譯。在3D圖形應用程式2310之編譯期間，高階著色器被編譯可為低階著色器。著色器指令2312可以以中間形式來提供，諸如由Vulkan API所使用之標準可攜式中間表示(SPIR)的版本。

【0347】使用者模式圖形驅動器2326可含有後端著色器編譯器2327，以將著色器指令2312轉換為硬體專用表示。當使用OpenGL API時，以GLSL高階語言之著色器指令2312被傳遞至使用者模式圖形驅動器2326以進行編譯。使用者模式圖形驅動器2326可使用作業系統內核模式功能2328以與內核模式圖形驅動器2329通訊。內核模式圖形驅

動器 2329 可以與圖形處理器 2332 通訊以分派命令和指令。

IP 核心實施方式

【0348】一或多個方面可以藉由儲存在機器可讀取媒體上之代表性碼來實現，該代表性碼表示及/或界定諸如處理器之積體電路內的邏輯。例如，機器可讀取媒體可包括其代表各種在處理器中的邏輯之指令。當由機器讀取時，指令可導致機器製造用於執行本文描述之技術的邏輯。被稱為「IP 核心」的此種表示是可以作為描述積體電路之結構之硬體模型儲存在有形的機器可讀取媒體上之積體電路之可重用的邏輯單元。硬體模型可以提供給各種消費者或製造設施，其在製造積體電路的製造機器上載入硬體模型。積體電路可被製造使得電路執行任何敘述在與本文描述的實施例相關的操作。

【0349】圖 24A 為根據實施例闡明可被用於製造積體電路以執行操作的 IP 核心開發系統 2400 的方塊圖。IP 核心開發系統 2400 可以用於產生模組化之可重複使用的設計，其可以併入更大的設計中或用於建構整個積體電路（例如，SOC 積體電路）。設計設施 2430 可以以高級編碼語言（例如，C/C++）產生 IP 核心設計的軟體模擬 2410。軟體模擬 2410 可被用來使用模擬模型 2412 設計、測試及驗證 IP 核心的行為。模擬模型 2412 可包括功能性、行為性及/或時序性模擬。暫存器轉移層級 (RTL) 設計 2415 可接著從模擬模型 2412 建立或合成。RTL 設計 2415 是對硬體暫存器之間

的數位訊號之流動進行建模的積體電路的行為的抽象，包括使用建模的數位訊號執行的相關聯的邏輯。除了RTL設計2415之外，還可以建立、設計或合成在邏輯層級或電晶體層級的較低層級設計。因此，初始設計和模擬之具體細節可以變化。

【0350】RTL設計2415或等效物可以由設計設施進一步合成為硬體模型2420，其可以是硬體描述語言(HDL)或物理設計資料的一些其他表示。HDL可被進一步模擬或測試以驗證IP核心設計。可以使用非揮發性記憶體2440來儲存IP核心設計以傳送到第三方製造設施2465(例如，硬碟、快閃記憶體或任何非揮發性儲存介質)。替代地，可以透過有線連接2450或無線連接2460來傳輸(例如，經由網際網路)IP核心設計。製造設施2465然後可以製造積體電路，其至少部分地基於IP核心設計。製造的積體電路可以被組態以執行根據本文描述的至少一實施例的操作。

【0351】圖24B繪示積體電路封裝總成2470的橫斷面側視圖。積體電路封裝總成2470繪示如本文所述之一或多個處理器或加速器裝置的實施方式。封裝總成2470包括連接至基材2480之硬體邏輯2472、2474的多個單元。邏輯2472、2474可以至少部分地以可組態邏輯或固定功能邏輯硬體來實施，並且可包括本文所述之處理器核心、圖形處理器、或其他加速器裝置中的任何一或多個部分。邏輯2472、2474的每個單元可以在半導體晶粒內實施，並經由互連結構2473與基材2480耦接。互連結構2473可經組態以

在邏輯 2472、2474 和基材 2480 之間路由電訊號，並且可以包括諸如但不限於凸塊或柱的互連。互連結構 2473 可經組態以路由電訊號，例如，與邏輯 2472、2474 的操作相關聯的輸入/輸出(I/O)訊號及/或電源或接地訊號。可選地，基材 2480 可以是基於環氧樹脂的疊層基材。基材 2480 亦可包括其他合適的基材類型。封裝總成 2470 可經由封裝互連 2483 連接至其他電性裝置。封裝互連 2483 可以耦接至基材 2480 的表面，以將電訊號路由至其他電性裝置，諸如主基板、其他晶片組、或多晶片模組。

【0352】邏輯 2472、2474 之單元可與組態以在邏輯 2472、2474 之間路由電訊號的電橋 2482 電性耦接。電橋 2482 以是提供用於電訊號之路徑的密集互連結構。電橋 2482 可包括電橋基材，其由玻璃或合適的半導體材料組成。電性路由特徵可經形成在電橋基材上，以提供介於邏輯 2472、2474 之間的晶片至晶片連接。

【0353】儘管繪示邏輯 2472、2474 的兩個單元和電橋 2482，但是本文描述的實施例可以在一或多個晶粒上包括更多或更少的邏輯單元。一或多個晶粒可藉由零或多個電橋，如當邏輯被包括在單一晶粒上時可排除電橋 2482。替代地，可藉由一或多個電橋連接多個晶粒或邏輯單元。額外地，多個邏輯單元、晶粒及電橋可以以其他可能的組態(包括三維組態)連接在一起。

【0354】圖 24C 繪示封裝總成 2490，其包括連接至基材 2480(例如，基底晶粒)之多個硬體邏輯小晶片單元。如

本文所述之圖形處理單元、並行處理器、及/或計算加速器可以由分別製造的各種矽小晶片組成。在此背景中，小晶片係至少部分的封裝積體電路，其包括可以與其他小晶片組裝成較大封裝之不同的邏輯單元。可以將具有不同IP核心邏輯的各種小晶片組裝到單一裝置中。額外地，小晶片可使用有源中介層技術整合至基底晶粒或基底小晶片。本文敘述的概念使得能夠在GPU內的不同形式的IP之間進行互連和通訊。IP核心可使用不同的製程技術製造並且可在製造期間組裝，從而避免將多個IP(尤其是在具有多個IP的大型SoC上)整合到同一製造製程的複雜性。啟用多種製程技術的使用可以縮短上市時間，並提供一種經濟高效的方式來創建多個產品SKU。額外地，分解後的IP更適合獨立供電，可以關閉給定工作負載中未使用的組件的電源，從而降低總體功耗。

【0355】在各種實施例中，封裝總成2490可以包括更少或更多數量之藉由結構2485或一或多個電橋2487互連的組件和小晶片。封裝總成2490內的小晶片可具有使用基材上晶圓上晶片堆疊的2.5D配置，其中多個晶粒並排堆疊在包括遍歷矽通孔(TSV)以將小晶片與基材2480耦接之矽中介層上，該基材包括至封裝互連2483的電性連接。

【0356】在一實施例中，矽中介層係有源中介層2489，其除了TSV之外還包括嵌入式邏輯。在這樣的實施例中，使用在有源中介層2489的頂部上的3D面對面晶粒堆疊來配置封裝總成2490內的小晶片。除了互連結構2485和

矽電橋 2487 外，有源中介層 2489 還可以包括 I/O 2491 的硬體邏輯、快取記憶體 2492 和其他硬體邏輯 2493。結構 2485 啟用有源中介層 2489 中各種邏輯小晶片 2472、2474 與邏輯 2491、2493 之間的通訊。結構 2485 可以是 NoC 互連，或者是在封裝總成的組件之間切換資料封裝的另一形式的封裝切換結構。對於複雜總成，結構 2485 可以是專用小晶片，以實現封裝總成 2490 的各種硬體邏輯之間的通訊。

【0357】有源中介層 2489 內的電橋結構 2487 可用來促進例如邏輯或 I/O 小晶片 2474 與記憶體小晶片 2475 之間的點對點互連。在一些實施方式中，電橋結構 2487 也可嵌入在基材 2480 內。

【0358】硬體邏輯小晶片可包括專用硬體邏輯小晶片 2472、邏輯或 I/O 小晶片 2474 及 / 或記憶體小晶片 2475。硬體邏輯小晶片 2472 及邏輯或 I/O 小晶片 2474 可以至少部分地以可組態邏輯或固定功能邏輯硬體來實施，並且可包括本文所述之處理器核心、圖形處理器、並行處理器、或其他加速器裝置中的任何一或多個部分。記憶體小晶片 2475 可為 DRAM (例如，GDDR、HBM) 記憶體或快取 (SRAM) 記憶體。有源中介層 2489 (或基材 2480) 內的快取記憶體 2492 可用封裝總成 2490 的全局快取、分散式全局快取的一部分、或用作結構 2485 的專用快取。

【0359】每一小晶片可製造為單獨的半導體晶粒並且與嵌入在基材 2480 內或與基材 2480 耦接的基底晶粒耦接。可經由互連結構 2473 來執行與基材 2480 的耦接。互連結構

2473可經組態以在每一小晶片與基材2480內的邏輯之間路由電訊號。互連結構2473可包括諸如但不限於凸塊或柱的互連。在一些實施例中，互連結構2473可經組態以路由電訊號，諸如(例如)輸入/輸出(I/O)訊號及/或與邏輯/I/O及記憶體小晶片之操作相關聯的電源或接地訊號。在一實施例中，額外的互連結構將有源中介層2489與基材2480耦接。

【0360】基材2480可為基於環氧樹脂的疊層基材，然而，不限於此，並且基材2480亦可包括其他合適類型的基材。封裝總成2490可經由封裝互連2483連接至其他電性裝置。封裝互連2483可以耦接至基材2480的表面，以將電訊號路由至其他電性裝置，諸如主基板、其他晶片組、或多晶片模組。

【0361】邏輯或I/O小晶片2474及記憶體小晶片2475可經由電橋2487電性耦接，電橋係組態以在邏輯或I/O小晶片2474和記憶體小晶片2475之間路由電訊號。電橋2487以是提供用於電訊號之路徑的密集互連結構。電橋2487可包括電橋基材，其由玻璃或合適的半導體材料組成。電性路由特徵可經形成在電橋基材上，以提供介於邏輯或I/O小晶片2474和記憶體小晶片2475之間的晶片至晶片連接。電橋2487亦可稱為矽電橋或互連電橋。例如，電橋2487係嵌入式多晶粒互連電橋(EMIB)。替代地，電橋2487可簡單地為從一個小晶片至另一小晶片的直接連接。

【0362】圖24D繪示根據一實施例之包括可互換小晶

片 2495 的封裝總成 2494。可互換小晶片 2495 可被組裝到一或多個基底小晶片 2496、2498 上的標準化槽中。基底小晶片 2496、2498 可經由電橋互連 2497 耦接，電橋互連 2497 可以類似於本文所述的其他電橋互連，並且可以是例如 EMIB。記憶體小晶片亦可經由電橋互連連接至邏輯或 I/O 小晶片。I/O 和邏輯小晶片可經由互連結構通訊。基底小晶片可各自以用於邏輯或 I/O 或記憶體/快取之一的標準化格式來支持一或多個槽。

【0363】可以將 SRAM 和電力傳輸電路製造到一或多個基底小晶片 2496、2498 中，其可以相對於堆疊在基底小晶片頂部的可互換小晶片 2495 使用不同的製程技術來製造它們。例如，可以使用較大的製程技術來製造基底小晶片 2496、2498，而可以使用較小的處理技術來製造可互換小晶片。可互換小晶片 2495 中的一或多個可以是記憶體(例如，DRAM)小晶片。可以基於使用封裝總成 2494 的產品的目標功率及/或效能為封裝總成 2494 選擇不同的記憶體密度。額外地，可以在組裝時根據產品的功率及/或效能來選擇具有不同數量功能單元類型的邏輯小晶片。額外地，可以將含有不同類型 IP 邏輯核心之小晶片插入可互換小晶片槽，從而實現可以混合和匹配不同技術 IP 區塊的混合處理器設計。

例示性系統單晶片積體電路

【0364】圖 25-26B 繪示可使用一或多個 IP 核心來製造

的例示性積體電路及相關聯的圖形處理器。除了所繪示的之外，可包括其他的邏輯和電路，包括額外的圖形處理器/核心、週邊介面控制器或通用處理器核心。具有與本文中任何其他圖式之元件相同或相似名稱之圖 25-26B 的元件可如本文其他地方所敘述的那樣、可以按照類似於本文所述的方式操作或起作用、可以包括相同的組件、並且可以鏈接到其他實體，但是不限於此。

【0365】圖 25 係繪示可使用一或多個 IP 核心來製造的例示性系統單晶片積體電路 2500 的方塊圖。例示性積體電路 2500 包括一或多個應用程式處理器 2505 (例如，CPU)、至少一圖形處理器 2510，其可為圖形處理器 1408、1508、2510 或本文所述之任何圖形處理器的變化，並且可取代所述之任何圖形處理器來使用。因此，本文結合圖形處理器的任何特徵之揭露亦揭露了與圖形處理器 2510 的相應組合，但不限於此。積體電路 2500 可額外地包括影像處理器 2515 及/或視訊處理器 2520，它們中的任何一個可以是來自相同或多個不同設計設施的模組化 IP 核心。積體電路 2500 可包括週邊或匯流排邏輯，包括 USB 控制器 2525、UART 控制器 2530、SPI/SDIO 控制器 2535、及 I²S/I²C 控制器 2540。額外地，積體電路可包括顯示裝置 2545，其耦接至一或多個高解析度多媒體介面 (HDMI) 控制器 2550 及行動產業處理器介面 (MIPI) 顯示介面 2555。可藉由包括快閃記憶體及快閃記憶體控制器的快閃記憶體子系統 2560 來提供儲存。記憶體介面可經由記憶體控制器 2565 來提供，以

對於 SDRAM 或 SRAM 記憶體裝置存取。一些積體電路額外包括嵌入式安全性引擎 2570。

【0366】圖 26A-26B 係根據文中所述之實施例繪示之在 SoC 內使用的例示性圖形處理器。圖形處理器可為圖形處理器 1408、1508、2510 或文中所述之任何其他圖形處理器的變化。可以使用圖形處理器取代圖形處理器 1408、1508、2510 或文中所述之任何其他圖形處理器。因此，與圖形處理器 1408、1508、2510 或文中所述之任何其他圖形處理器組合之任何特徵的揭露亦揭露了與圖 26A-26B 之圖形處理器的相應組合，但不限於此。圖 26A 繪示根據一個實施例之可使用一或多個 IP 核心來製造的系統單晶片積體電路的例示性圖形處理器 2610。圖 26B 繪示根據一個實施例可使用一或多個 IP 核心來製造的系統單晶片積體電路之額外的例示性圖形處理器 2640。圖 26A 之圖形處理器 2610 係低功率圖形處理器核心的實例。圖 26B 之圖形處理器 2640 係高效能圖形處理器核心的實例。例如，如本段落的開頭所提到的，圖形處理器 2610 和圖形處理器 2640 中的每一個可以是圖 25 的圖形處理器 2510 的變化。

【0367】如圖 26A 所示，圖形處理器 2610 包括頂點處理器 2605 及一或多個片段處理器 2615A-2615N (例如 2615A、2615B、2615C、2615D 到 2615N-1 和 2615N)。圖形處理器 2610 可經由個別的邏輯來執行不同的著色器程式，以致使頂點處理器 2605 被優化以執行針對頂點著色器程式之操作，而一或多個片段處理器 2615A-2615N 係執行針對

片段或像素著色器程式之片段(例如，像素)著色操作。頂點處理器 2605 係執行 3D 圖形管線之頂點處理階段並產生基元及頂點資料。片段處理器 2615A-2615N 係使用由頂點處理器 2605 所產生的基元及頂點資料以產生被顯示在顯示裝置上的框緩衝器。片段處理器 2615A-2615N 可被優化以執行如針對 OpenGL API 中所提供的片段著色器程式，其可被用來執行如針對 Direct 3D API 中所提供的像素著色器程式之類似操作。

【0368】圖形處理器 2610 額外地包括一或多個記憶體管理單元 (MMU) 2620A-2620B、快取 2625A-2625B 及電路互連 2630A-2630B。一或多個 MMU 2620A-2620B 為圖形處理器 2610 (包括頂點處理器 2605 及 / 或片段處理器 2615A-2615N) 提供虛擬至實體位址映射，除了儲存在一或多個快取 2625A-2625B 中的頂點或影像 / 紋理資料之外，其可以參考儲存在記憶體中的頂點或影像 / 紋理資料。一或多個 MMU 2620A-2620B 可以與系統內的其他 MMU 同步，包括與圖 25 的一或多個應用程式處理器 2505、影像處理器 2515 及 / 或視訊處理器 2520 相關聯的一或多個 MMU，使得每個處理器 2505-2520 可以參與共享或統一虛擬記憶體系統。圖形處理器 2610 的組件可對應於本文敘述的其他圖形處理器的組件。一或多個 MMU 2620A-2620B 可對應於圖 2C 的 MMU 245。頂點處理器 2605 和片段處理器 2615A-2615N 可以與圖形多處理器 234 相對應。根據實施例，一或多個電路互連 2630A-2630B 經由 SoC 的內部匯流排或經由直接連

接致使了圖形處理器 2610 與 SoC 內的其他 IP 核心介接。一或多個電路互連 2630A-2630B 可對應於圖 2C 的資料交叉開關 240。可以在圖形處理器 2610 的類似組件與本文敘述的各種圖形處理器架構之間找到進一步的對應關係。

【0369】如圖 26B 所示，圖形處理器 2640 包括圖 26A 的圖形處理器 2610 的一或多個 MMU 2620A-2620B、快取 2625A-2625B 及電路互連 2630A-2630B。圖形處理器 2640 包括一或多個著色器核心 2655A-2655N(例如，2655A、2655B、2655C、2655D、2655E、2655F 至 2655N-1 及 2655N)，其提供統一的著色器核心架構，其中單一核心或核心類型可執行所有類型的可編程著色器碼(包括著色器程式碼)以實現頂點著色器、片段著色器及/或計算著色器。所存在之著色器核心的精確數目可在實施例及實施方式之間變化。額外地，圖形處理器 2640 包括核心間任務管理器 2645，其充當用以將執行緒分派給一或多個著色器核心 2655A-2655N 的執行緒分派器以及用以加速基於渲染之鋪磚操作的鋪磚單元 2658，其中在影像空間中細分場景的渲染操作，例如以利用場景內的局部空間一致性或優化內部快取的使用。著色器核心 2655A-2655N 可以對應於例如圖 2D 中的圖形多處理器 234，或分別對應於圖 3A 和 3B 的圖形多處理器 325、350、或圖 3C 的多核心群組 365A。

稀疏資料的脈衝式算術

【0370】稀疏矩陣是相對於矩陣中元素總數而言，包

含少量非零元素的矩陣。矩陣的稀疏度可以藉由零值元素的數量除以元素的總數來指明。可以藉由壓縮資料及/或將資料編碼為稀疏編碼格式來減少稀疏矩陣消耗的儲存量。藉由旁路零值矩陣元素上的乘法操作，可以加速使用稀疏矩陣執行的處理操作。還可以對主要是任何單值或可容易壓縮的矩陣資料進行優化。

【0371】文中所述之實施例包括軟體、韌體、及硬體邏輯，其提供經由脈動處理單元或其他矩陣及/或張量加速器單元對稀疏資料執行算術的技術。一個實施例提供當使用稀疏資料時優化對脈動陣列的訓練和推理的技術。一個實施例提供了在執行稀疏計算操作時使用解壓縮資訊的技術。一個實施例提供了一種數值轉換，以轉換稀疏資料並使得能夠基於轉換後的資料進行訓練。一個實施例提供了直接對壓縮資料執行脈衝式算術的技術。

具有張量加速邏輯的 GPGPU

【0372】圖 27 係根據實施例之資料處理系統 2700 的方塊圖。資料處理系統 2700 係具有處理器 2702、統一記憶體 2710、及包括機器學習加速邏輯之 GPGPU 2720 的異質處理系統。處理器 2702 及 GPGPU 2720 可以是本文所述的任何處理器和 GPGPU/並行處理器。處理器 2702 可執行用於儲存在系統記憶體 2712 中的編譯器 2715 的指令。編譯器 2715 在處理器 2702 上執行，以將來源碼 2714A 編譯為已編譯碼 2714B。已編譯碼 2714B 可包括由處理器 2702 執行的

指令及/或由 GPGPU 2720執行的指令。在編譯期間，編譯器 2715可執行插入元資料的操作，包括關於已編譯碼 2714B中存在的資料並行度的提示及/或關於與將基於已編譯碼 2714B分派的執行緒相關聯的資料位置的提示。編譯器 2715可包括執行這種操作所需的資訊，或者可以在運行時庫(runtime library)2716的幫助下執行這些操作。運行時庫 2716亦可幫助編譯器 2715進行來源碼 2714A的編譯，並且亦可包括在運行時與已編譯碼 2714B鏈接的指令，以便於在 GPGPU 2720上執行已編譯指令。

【 0373 】 統一記憶體 2710表示可由處理器 2702和 GPGPU 2720存取的統一位址空間。統一記憶體可包括系統記憶體 2712以及 GPGPU記憶體 2718。GPGPU記憶體 2718係在 GPGPU 2720之位址空間內的記憶體並且可包括一些或全部的系統記憶體 2712。在一實施例中，GPGPU記憶體 2718亦可包括專用於由 GPGPU 2720專用的任何記憶體的至少一部分。在一實施例中，可以將儲存在系統記憶體 2712中的以編譯碼 2714B映射至 GPGPU記憶體 2718，以供 GPGPU 2720存取。

【 0374 】 GPGPU 2720包括多個計算區塊 2724A-2724N，其可以包括本文敘述之各種處理資源中的一或多個。處理資源可以是或包括各種不同的計算資源，諸如執行單元、計算單元、串流多處理器、圖形多處理器或多核心群組。在一實施例中，GPGPU 2720還包括張量加速器 2723，其可以包括一或多個特殊功能計算單元，其被設計

為加速矩陣操作的子集(例如，內積等)。張量加速器 2723 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 2723 內的邏輯組件可分佈在多個計算區塊 2724A-2724N 的處理資源上。

【0375】 GPGPU 2720 亦可包括可以由計算區塊 2724A-2724N 和張量加速器 2723 共享的一組資源，包括但不限於一組暫存器 2725、功率和效能模組 2726、及快取 2727。在一實施例中，暫存器 2725 包括直接和間接可存取暫存器，其中直接可存取暫存器被優化以供張量加速器 2723 使用。功率和效能模組 2726 可被組態以調整計算區塊 2724A-2724N 的功率遞送和時脈頻率，以給計算區塊 2724A-2724N 內的閘閏置組件供電。在各種實施例中，快取 2727 可包括指令快取及/或較低層級資料快取。

【0376】 GPGPU 2720 可額外地包括 L3 資料快取 2730，其可用於快取張量加速器 2723 及/或計算區塊 2724A-2724N 內的計算元件從統一記憶體 2710 存取的資料。在一實施例中，L3 資料快取 2730 包括由計算區塊 2724A-2724N 和張量加速器 2723 內的計算元件共享的共享本地記憶體 2732。

【0377】 在一實施例中，GPGPU 2720 包括指令處置邏輯，諸如提取和解碼單元 2721 以及排程器控制器 2722。提取和解碼單元 2721 包括提取單元和解碼單元，用以提取和解碼由計算區塊 2724A-2724N 或張量加速器 2723 中的一或多個執行的指令。可經由排程器控制器 2722 將指令排程

至計算區塊 2724A-2724N 內適當功能單元或張量加速器。在一實施例中，排程器控制器 2722 係可組態以執行先進排程操作的 ASIC。在一實施例中，排程器控制器 2722 係能夠執行從韌體模塊載入的排程器指令的微控制器或每指令低能耗的處理核心。

【0378】在一實施例中，可以將由計算區塊 2724A-2724N 執行的一些功能直接排程到張量加速器 2723 或將其卸載。在各種實施例中，張量加速器 2723 包括組態以有效地執行矩陣計算操作的處理元件邏輯，諸如 3D 圖形或計算著色器程式所使用的乘法和加法操作以及內積操作。在一實施例中，張量加速器 2723 可經組態以加速由機器學習框架使用的操作。在一實施例中，張量加速器 2723 係特定應用積體電路，其被明確組態以執行一組特定的並行矩陣乘法及/或加法操作。在一實施例中，張量加速器 2723 係現場可程式閘陣列 (FPGA)，其提供可在工作負載之間更新的固定功能邏輯。可以由張量加速器 2723 執行的矩陣操作組相對於可以由計算區塊 2724A-2724N 執行的操作可能受到限制。然而，張量加速器 2723 可以相對於計算區塊 2724A-2724N 以明顯更高的通量執行那些操作。

【0379】圖 28 繪示根據實施例之由指令管線 2800 執行的矩陣操作 2805。指令管線 2800 可經組態以執行矩陣操作 2805，諸如但不限於內積操作。兩個向量的內積係一純量值，其等於向量的相應分量之乘積的和。可以以如下所述的方程式 (1) 計算內積。

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

【0380】內積可用於卷積類神經網路(CNN)的卷積操作。圖28繪示使用包括內積操作的矩陣操作2805的二維(2D)卷積。儘管繪示2D卷積，但是可以使用N維濾波器在N維體積上執行N維卷積。接收欄位磚片2802突顯輸入量緩衝器2804中的一部分輸入量。輸入量緩衝器可儲存在記憶體2830中。可在接收欄位磚片2802內的資料與卷積濾波器之間執行點矩陣操作2805，以在輸出緩衝器2806內產生資料點，其亦可儲存在記憶體2830中。記憶體2830可以是本文所述的任何記憶體，包括系統記憶體2712、GPGPU記憶體2718或如圖27中的一或多個快取記憶體2727、2730。

【0381】輸出緩衝器2806內資料點的組合表示由卷積操作產生的激勵圖(activation map)。激勵圖內的每一點係藉由在輸入量緩衝器2804滑動接收場磚片產生的。可將激勵圖資料輸入至激勵函數以判定輸出激勵值。在一實施例中，輸入量緩衝器2804之卷積可在框架內界定為高階矩陣操作2905。可經由基元操作(諸如，基本線性代數子程式(BLAS)操作)執行高階矩陣操作。基元操作可經由由指令管線2800執行的硬體指令加速。

【0382】用來加速硬體指令的指令管線2800可包括可提取和解碼硬體指令的指令提取和解碼單元2721及排程器控制器2722，其可將已編碼指令排程至計算區塊2724A-2724N及/或張量加速器2723中的一或多個處理資源。在一

實施例中，可將硬體指令排程至計算區塊 2724A-2724N，並將其卸載至張量加速器 2723。可以將一或多個硬體指令和用以執行矩陣操作 2805 的相關聯資料儲存在記憶體 2830 中。硬體指令之輸出亦可儲存在記憶體 2830 中。

【0383】在一實施例中，張量加速器 2723 可執行一或多個指令已使用積體脈動張量陣列 2808 (DP 邏輯) 來執行矩陣操作 2805。脈動張量陣列 2808 可包括可編程和固定功能硬體的組合，其可組態以執行內積操作。雖然計算區塊 2724A-2724N 內的功能單元也可以組態以執行內積操作，但脈動張量陣列 2808 可組態以相對於計算區塊 2724A-2724N 顯著更高的通量執行內積操作的有限子集。

【0384】圖 29A-29B 繪示根據一些實施例之基於硬體的脈動張量陣列 2808 的細節。圖 29A 繪示多個功能單元的網格，其可組態為在單一時脈循環內執行多個內積操作。圖 29B 繪示單一例示性功能單元。

【0385】如圖 29A 所示，在一實施例中，脈動張量陣列 2808 係組態以使用各種功能單元執行一組並行內積操作。內積可以以「脈動」方式執行，其中 SIMD 資料係跨多個功能單元層進行泵送。脈動張量陣列 2808 係配置在網格中功能單元的集合。功能單元的網格按步調工作，並經過優化以執行乘法累加操作。將由脈動張量陣列 2808 操作的矩陣分為子矩陣，這些子矩陣被泵送到功能單元的網格上。

【0386】在一實施例中，脈動張量陣列 2808 可使用可

組態脈動深度來處理可組態數量的 SIMD 資料通道。對於給定的指令，可以選擇 SIMD 寬度及脈動深度來處理一組來源資料。脈動深度界定將被用來處理指令之硬體邏輯的脈動層的數量。脈動層係一組具有可變 SIMD 寬度的乘法和加法器邏輯單元，其中脈動層可接收初始累加器值作為輸入，並產生內積值以輸出至後續脈動層或輸出暫存器。

【0387】在一些實施例中，可處理三個來源，其中每一來源可以是向量暫存器或立即數。在一實施例中，來源 2900(SRC0)可為一或多個初始累加器值，其可以是單一值或累加器值的向量。初始累加器值將添加到第一脈動層內每一功能單元計算出的第一組內積。對於給定的 SIMD 通道，由功能性單元計算出的內積可被提供至下一個脈動層。可基於來源 2901(SRC1)和來源 2902(SRC2)計算內積，其係可含有封裝資料之一或多個通道向量暫存器，每一通道含有四元素向量。在一實施例中，每一通道係 32 位元寬並提供四個 8 位元向量元素。一些實施例係可組態以從具有 8 位元元素、4 位元元素及/或 2 位元元素的輸入向量來計算內積。在一實施例中，可以使用支持的元素大小的任意組合(例如，8 位元 $\times$ 2 位元、8 位元 $\times$ 4 位元、4 位元 $\times$ 4 位元等)來執行混合精度操作。在一實施例中，脈動張量陣列 2808 係組態用於整數計算，儘管在某些實施例中可配置自動共定點操作。儘管本文敘述的指令是四元素內積，但是在一些實施例中，脈動張量陣列 2808 亦可組態為支持在每

個向量上不同數量之元素上的浮點內積計算。

【0388】在一實施例中，可以將四元素向量的多個通道封裝到具有各種寬度(例如 64 位元、128 位元、256 位元、512 位元等)的單一向量暫存器中。可以經由脈動張量陣列 2808 為經由來源 2901 和來源 2902 提供的向量元素的多個通道計算同步內積。可以基於為內積計算選擇的執行大小和脈動深度來組態要處理的向量元素的通道數量。在一實施例中，可以使用脈動張量陣列 2808 的多個循環來計算比指定的執行大小及/或脈動深度更寬的來源向量。

【0389】在給定的時脈循環內可以執行的計算數量可基於 SIMD 通道和脈動層的數量而變化。如所示之脈動張量陣列 2808 可以使用 4 的脈動深度在每一 SIMD 通道通道上執行十六個內積。如果為八個 SIMD 通道組態，則該邏輯可以在給定的周期內執行 128 個八位元整數(INT8)內積。如果為八個 SIMD 通道組態且脈動深度為八，則每一通道可以執行 32 個八位元整數(INT8)內積和 256 個內積。這些特定數量的操作是一個例示性實施例，而其他實施例的通道有所不同。此外，如果資料類型不同，則將基於不同的資料類型縮放操作數量。

【0390】在每一功能單元上，經由乘法器和加法器邏輯計算內積，並將內積加到累加器值上。所得資料可以輸出至目的地暫存器或提供給下一個脈動層的累加器。功能單元 2912 的細節在圖 29B 中繪示。

【0391】如圖 29B 所示，功能單元 2912 可以包括一組

輸入資料緩衝器 2904、2906 和累加器 2922，它們可以各自接受輸入資料。在一實施例中，資料緩衝器 2906 可接受來源 2902(SRC2)，其可為封裝的輸入資料的向量。輸入資料緩衝器 2904 可接受來源 2901(SRC1)，其亦可為封裝的輸入資料的向量。累加器 2922 可以接受為功能單元 2912 提供初始累加器值的來源 2900(SRC0)。將初始累加器值添加至從來源 2901 和來源 2902 的元素計算出的內積。內積係使用組乘法器 2923A-2329D 和加法器 2924 經由來源向量的逐元素乘法來計算。乘法器 2932A-2923D 用以計算一組乘積。由加法器 2924 計算該組乘積的總和。所述總和可以與經由來源 2900 提供的任何起始值累加(例如，相加)。在一實施例中，可以將所述累加值作為輸入值 2926 提供給下一個累加器，該下一個累加器可以位於隨後脈動層中。在一實施例中，來源 2901 可包括輸入資料的多個通道。可以將來源 2901 的額外通道作為 SRC1 輸入中繼到額外 SIMD 通道 2928。在一實施例中，來源 2902 可包括輸入資料的多個通道。來源 2902 的額外通道可以用作 SRC2 輸入資料，以到達其他脈動深度內的邏輯單元。在一實施例中，來源 2900 可選地包括多個通道，在額外的功能單元內提供額外通道作為累加器的輸入。在一實施例中，來源 2900 可以是被添加到初始脈動層之每一功能單元中的每個累加器的單值。

【0392】圖 30 繪示根據一個實施例之包括具有稀疏優化的脈動張量陣列的處理資源 3000。處理資源 3000 包括與

執行單元 1900 類似的組件，以及具有稀疏優化的張量加速器 3012。可以在並行計算加速器、圖形處理引擎及/或類神經網路加速器中使用處理資源 3000 的變化。張量加速器 3012 可以包括如圖 28 中的脈動張量陣列 2808 或如圖 19 中的脈動陣列 1912，並且具有相似的功能。亦可以使用其他矩陣加速架構。張量加速器 3012 的稀疏優化在下面詳細描述。

當使用稀疏資料時對脈動陣列的優化訓練和推理

【0393】 一個實施例提供當使用稀疏資料時優化對脈動陣列的訓練和推理的技術。如果由脈動張量陣列 2808 處理的子矩陣完全為零，則可將子矩陣的尺寸值設置為零，並且脈動張量陣列 2808 可以取決於要執行的操作旁路與子矩陣相關的一或多個計算相位。在矩陣資料的預處理期間，可識別零子矩陣，並且可以產生該矩陣的子矩陣圖以指示哪些子矩陣僅包含零值。在一實施例中，也可旁路僅包括一個非零值的子矩陣。

【0394】 圖 31A-31B 繪示根據實施例之旁路零值子矩陣的系統。如圖 31A 所示，矩陣 3102 和矩陣 3104 是其中一或多個子矩陣僅含有零值的矩陣。處理邏輯可以產生用於矩陣 3102 的子矩陣圖 3112 和用於矩陣 3104 的子矩陣圖 3114，以指示子矩陣是否僅含有零值。可以使用多種技術來產生子圖，包括對每一子矩陣執行與零的按位比較。子矩陣圖可以由在通用處理邏輯(例如，CPU)上執行的框架

或驅動器邏輯來產生，或者可以由處理資源內的專用硬體邏輯來產生。

【0395】在一實施例中，還可以旁路包括單一非零值的單一非零值子矩陣 3105，因為可以使用 ALU 而不是脈動張量陣列來計算矩陣操作的結果。

【0396】如圖 31B 所示，記憶體 3120 可儲存矩陣 3102 和矩陣 3104。脈動張量陣列 2808 可包括矩陣 A 載入單元 3126、矩陣 B 載入單元 3122、矩陣 A 饋送單元 3128、及矩陣 B 饋送單元 3124。矩陣 3102 可以作為矩陣 B 載入和饋送，而矩陣 3104 可作為矩陣 A 載入和饋送。矩陣 A 和矩陣 B 的子矩陣可以載入並透過功能單元 3130 饋送，這些功能單元用作脈動張量陣列 2808 的處理元素。

【0397】在一實施例中，負載 B 濾波器 3127 和負載 A 濾波器 2127 可包括緩衝器，以儲存矩陣 3102 的子矩陣圖 3112 和矩陣 3104 的子矩陣圖 3114。負載 B 濾波器 3121 可以藉由矩陣 B 負載單元 3122 旁路零值子矩陣的負載。負載 A 濾波器 3127 可以藉由矩陣 A 負載單元 3126 旁路零值子矩陣的負載。功能單元 3130 可以處理未被旁路的子矩陣。取決於要由脈動張量陣列 2808 執行的操作，其中子矩陣之一為零，可以旁路整個操作。當子矩陣包括單一非零值子矩陣時，與要執行的操作相關的子矩陣可以旁路脈動張量陣列 2808，並且該操作可以由 ALU 執行。

【0398】圖 32 繪示計算資源 3010，其包括旁路脈動張量陣列 2808 以用於在子矩陣上操作的邏輯，該子矩陣包括

單一非零值。可以將儲存在記憶體 3120 中的矩陣 3102 和矩陣 3104 的元素載入到計算資源 3010 的暫存器檔 3006 中。張量加速器 3012 可包括用以向 ALU 3011 傳輸子矩陣旁路訊息 3202 的邏輯，其指示儲存了被旁路的子矩陣資料 3204 和要執行的旁路操作的暫存器。然後，ALU 3011 可以使用向量處理邏輯執行旁路操作。可以與由張量加速器 3012 執行的非旁路操作並行執行旁路操作的處理。

當執行稀疏計算操作使用解壓縮資訊

【0399】圖 33 根據一實施例繪示組態以能夠將類神經網路資料壓縮傳輸至並行計算處理器或通用圖形處理單元上的處理資源的計算架構 3300。在一實施例中，計算架構 3300 包括計算區塊 3302 及經由 DMA 控制器 3306 耦接至記憶體 3308 的硬體高速暫存緩衝器 3304。記憶體 3308 可為資料處理系統的主記憶體或系統記憶體。計算區塊 3302 包括如本文所述之一組處理資源，並且可類似於圖 27 中的任何計算區塊 2724A-2724N。高速暫存緩衝器 3304 可為高速晶載記憶體，諸如晶載靜態隨機存取記憶體 (SRAM)。在一實施例中，高速暫存緩衝器 3304 被優化以儲存特徵區塊單元或內核區塊單元，以用於由計算區塊 3302 執行的類神經網路操作。

【0400】在一實施例中，解碼器 3312 可以是硬體解碼器邏輯，其被整合到計算區塊 3302 中以使得能夠在計算架構上進行類神經網路資料的壓縮傳輸。例如，當處理 CNN

時，計算區塊 3302 可以以未壓縮格式在高速暫存緩衝器 3304 中產生輸出特徵圖 (OFM) 資料。整合在 DMA 控制器 3306 中以使能夠寫入輸出特徵的編碼器 3316 以壓縮格式將資料映射至記憶體 3308。當一層的 OFM 成為下一層的輸入特徵圖 (IFM) 時，這些 IFM 為壓縮資料 3314 從記憶體 3308 中讀取，並儲存在高速暫存緩衝器 3304 中。解碼器 3312 可以使計算區塊 3302 能夠讀入壓縮資料 3314，而不需要對資料進行解碼。替代地，具有編碼和解碼邏輯的編碼解碼器單元可整合至 DMA 控制器 3306 中，從而使壓縮資料可以由 DMA 控制器 3306 傳輸和讀取。然後，可以由 DMA 控制器 3306 對特徵圖資料進行解壓縮，並以未壓縮的格式將其寫入暫存緩衝區 3304，以供計算區塊 3302 讀取。

【0401】在本文所述之實施例中，可以基於要編碼的資料的統計資訊來改變用於內核和特徵資料的特定編碼格式。類神經網路特徵圖資料的分析指示許多特徵圖可能非常稀疏。類神經網路內核資料的分析指示儘管內核資料不像特徵圖資料那樣稀疏，但是內核資料中的許多值都是重複的。內核資料的動態範圍相對較低，其指示原資料分配的位元比儲存係數所需的位元更多。使用各種編碼技術，可以藉由選擇各種編碼技術以無損方式將特徵圖和內核資料壓縮多達 80%。

【0402】可以使用多種編碼技術來編碼(例如，壓縮)與類神經網路相關資料，例如但不限於唯一絕對值(UAV)表編碼、顯著圖(SM)編碼、表編碼(TE)、唯一值坐標

(UVC)編碼和均值編碼(ME)。編碼資料的元資料指示用於資料的編碼格式的類型。在一實施例中，可以為諸如內核資料或特徵資料之類的特定類型的資料選擇特定的編碼格式。在一實施例中，在編碼之前對資料執行統計分析，以使得能夠為每個資料區塊選擇適當的編碼器。

【0403】在一實施例中，在SM編碼期間產生的資料可以用於促進脈動張量陣列內的子矩陣旁路。在SM編碼模式中，僅對區塊中的非零值進行編碼。標頭指示樣本區塊中非零值的數量，其後是指示該區塊內非零值的映射的有效圖。然後，按照串流中出現的順序對樣本的非零值進行編碼。

【0404】圖34A-34B繪示根據實施例之用於稀疏類神經網路資料的顯著圖編碼。圖34A繪示用於SM編碼的編碼佈局3420。圖34B繪示例示性位元串流顯著圖編碼的位元串流的解碼。

【0405】如圖34A所示，編碼佈局3420使稀疏類神經網路資料的樣本區塊能夠以減少位元格式編碼，該減少的位元格式減少了在處理與資料相關的類神經網路時需要傳輸或儲存的資料量。在所示編碼佈局3420中，在標頭3422中指示樣本區塊中非零值的數量，其後是指示該區塊內非零值的映射的有效圖3424。樣本的非零值3426在串流中按照出現的順序進行編碼。

【0406】如圖34B所示，可以基於顯著圖來對編碼的位元流資料進行解碼。在一實施例中，SM編碼模式資料

從兩個位元的標頭之後的第三個位元組開始顯示，其中 SM 編碼的存在由已編資料之串流開頭處的位元串流標頭(未圖示)的第一位元指示。標頭 3422 中指示樣本區塊中非零值的數量。在一實施例中，編碼格式可在每個樣本區塊的基礎上改變，因此標頭 3422 還可以包括指示為即將到來的樣本區塊啟用 SM 編碼的元資料。顯著圖 3424 指示樣本區塊內的非零值的圖，以及與每個值相關聯的一位元條目。然後，按照串流中出現的順序對樣本的非零值 3426 進行編碼。將顯著圖資料解碼為例示性解碼的位元串流 3430，解碼器邏輯可以將輸出資料緩衝器 3410 的至少一部分初始為零。然後，解碼器邏輯可以參考顯著圖 3424 來判定位元流中哪個值不為零。非零值可以被順序解封裝並且被寫入到由顯著圖指示的輸出資料緩衝器 3410 中的位置。例如，顯著圖 3424 中的零值 (0b0) 指示對應的解碼值為零。顯著圖 3424 中的一值 (0b1) 指示對應的解碼值對應於編碼位元串流中的非零值 3426 中的下一個連續條目。

【0407】圖 35 繪示使用顯著圖資料來促進對脈動張量陣列的元件的旁路或時脈/功率門控。在一實施例中，計算區塊 3302 內的解碼器 3312 可以為使用顯著圖編碼編碼的解壓縮的資料 3508 提供顯著圖 3506。在一實施例中，解碼器可以為使用不同的(非顯著圖)編碼進行編碼的資料產生顯著圖 3506。可以由解碼器 3312 將顯著圖 3506 和解壓縮的資料 3508 提供給計算區塊 3404 內的處理資源 3000。脈動張量陣列 2808 內的旁路/閘邏輯 3510 可以基於由顯著圖 3506

指示的稀疏度及/或由顯著圖 3506 指明的特定零和非零元素判定旁路解壓縮的資料 3508 內的子矩陣及/或對脈動張量陣列內的某些處理元素進行時脈門控。

使用數值變換轉換稀疏資料

【0408】在一實施例中，訓練被擴展為包括權重的變換以壓縮同一區域中非零係數以減少記憶體頻寬的消耗。訓練系統可以測試幾種不同的變換，諸如離散餘弦變換 (DCT)、離散符號變換 (DST)、位元轉換變換、位旋轉變換、或其他類型的變換 (例如，離散傅立葉變換 (DFT)，等等。) 訓練系統可選擇在壓縮訓練資料時導致非零係數最接近的叢集及/或最高壓縮率的變換。在一實施例中，訓練系統可選擇將不規則稀疏轉換為結構化或區塊稀疏性的變換。所選的轉換模式在經訓練模型內用訊號發送。經變換訓練模型將提供給推理部署。推理系統應用適當的逆變換。

【0409】圖 36 繪示根據一實施例之用於將類神經網路資料轉換為緊湊的非零係數的系統 3600。系統 3600 包括訓練系統 3602、變換系統 3604、逆變換系統 3608 和推理系統 3610。訓練系統 3602 和變換系統 3604 可駐留在一個計算裝置或多計算裝置之叢集上。逆變換系統 3608 和推理系統 3610 可以駐留在單獨的計算裝置或計算裝置的叢集上。系統 3600 的組件還可以駐留在執行訓練和推理操作兩者的單個計算系統或計算叢集上，諸如執行基於雲端的訓練和推

理的基於資料中心的計算系統。

【0410】在要進行模型再訓練或更新的地方，也可以使用單一計算系統或計算叢集。經訓練模型的權重可以以變換和壓縮狀態儲存。當要對模型進行重新訓練或更新時，可以對權重進行解壓縮，可以對這些權重應用逆變換，並且可以執行重新訓練或更新以產生更新的權重集。然後可以對變換後的權重集進行重新變換和重新壓縮，以儲存或部署到推理系統。

【0411】在一實施例中，訓練系統3602可以包括如本文所述的張量加速器(例如，圖27中的張量加速器2723)，作為在類神經網路上執行的訓練操作的結果，可以產生權重矩陣3603。權重矩陣可以由變換系統3604處理。變換系統3604可以判定應用於權重矩陣3603的變換類型。然後，變換系統可以輸出一組變換後的權重3607，並將變換類型3605應用於這些變換後的權重。然後可以對這組變換的權重3607進行壓縮或編碼，以進行儲存或傳輸。執行該變換，以使一組變換後的權重3607壓縮到比權重原始矩陣3603更高的壓縮率。一組轉換權重3607可以壓縮到較高壓縮率可能是由於存在於一組轉換權重3607中的較高稀疏度或由於從不規則稀疏性到結構化稀疏性或區塊稀疏性的轉換。結構化或區塊稀疏性及/或更高度度的稀疏性還可以提高稀疏變換矩陣可以以稀疏編碼格式(例如，坐標列表編碼(COO)、壓縮稀疏列(CSR)、壓縮稀疏行(CSC)等)。在一實施例中，可將變換後的權重的稀疏矩陣都以稀疏編碼

格式編碼並壓縮成壓縮編碼格式。

【0412】可以儲存變換後的權重 3607 和變換類型 3605，以供組態以執行推理操作的計算系統使用。在壓縮變換後的權重 3607 的情況下，可以在逆變換系統 3608 應用逆變換之前對變換後的權重進行解壓縮。逆變換系統 3608 可以基於指示的變換類型 3605 對變換後的權重 3607 執行逆變換操作。逆變換系統 3608 可以將權重矩陣 3609 輸出到推理系統 3610，以基於輸入資料執行推理操作。

【0413】圖 37A-37B 繪示對稀疏類神經網路資料執行變換和逆變換的邏輯單元。圖 37A 繪示變換系統 3604 的邏輯單元。圖 37B 繪示逆變換系統 3608 的邏輯單元。變換系統 3604 和逆變換系統 3608 可各自包括內部記憶體及/或與其中包括變換系統 3604 和逆變換系統 3608 的計算裝置或計算系統的記憶體耦接。

【0414】如圖 37A 所示，變換系統 3604 可以包括記憶體 3701、變換選擇器 3702 和一組變換邏輯單元 3704A-3704N。該組變換邏輯單元 3704A-3704N 可以包括硬體或軟體邏輯，以執行例如 DCT、DST、位元翻轉、位元旋轉或其他變換(例如，DFT 等)。對於 DCT 變換，使用無損變換。然而，在一些組態中，如果模型的準確性沒有受到顯著影響，則可以應用有損 DCT。在一個實施例中，一組變換邏輯單元 3704A-3704N 被實施為可以由本文敘述的 GPGPU 執行的計算著色器或 CUDA 內核的形式的 GPGPU 可執行程式碼。

【0415】變換測試邏輯 3703 可將權重矩陣 3603 的代表性樣本載入到記憶體 3701 中並對樣本進行測試，以判定在壓縮訓練資料時產生最高稀疏度，最接近的非零係數叢集及/或最高壓縮率的變換。在一實施例中，變換測試邏輯 3703 可以分析變換測試資料並開發與測試變換資料內的非零值之間的距離有關的度量。在一實施例中，變換測試邏輯 3703 還可分析測試變換資料以判定關於變換資料的潛在壓縮率、稀疏度或稀疏模式的度量。在比較變換時將壓縮率用作品質度量的實施例中，變換測試邏輯 3703 可以包括對變換後的資料執行壓縮測試的邏輯。變換測試邏輯還可以判定是否不存在將產生預期結果的變換。在這種情況下，取決於系統組態，可以旁路轉換。

【0416】一旦判定了最佳變換，則變換測試邏輯 3703 然後可以向變換選擇器 3702 發送關於使用變換邏輯單元 3704A-3704N 的集合中的哪個單元來將權重矩陣變換為變換後的權重 3607 的推薦。在一實施例中，變換測試邏輯 3703 可以被組態以產生變換推薦而不將變換應用於矩陣資料。在這樣的實施例中，變換測試邏輯 3703 可以提供與數值變換測試相關聯的度量以及變換推薦。此外，在這樣的實施例中，當啟用變換時，變換選擇器 3702 還可以考慮由變換測試邏輯 3703 產生的測試度量。然後，在某些情況下，變換選擇器 3702 可以選擇除推薦變換之外的變換。例如，如果度量指示例如壓縮率的改善低於臨限，則可以旁路變換。該臨限可基於當前處理器狀態(例如，熱負載、

可用頻寬等)而變化。在一實施例中，如果指示變換系統 3604 應用特定變換，則旁路變換測試邏輯 3703。

【0417】在一實施例中，可以將不同的變換用於權重矩陣的子矩陣。例如，如果變換測試邏輯判定如果使用第一變換進行變換，則權重矩陣的第一部分將具有最高的壓縮率，而如果使用第二變換進行變換，則權重矩陣的第二部分將具有最高的壓縮率，可以將第一變換應用於第一部分，而將第二變換應用於第二部分。在將多個不同的變換用於權重矩陣的情況下，變換類型 3605 可以是指示哪種變換類型用於對應的權重子矩陣的矩陣或位元欄。在一些實施例中，整體上還可以將多個變換應用於相同的子矩陣或權重矩陣。

【0418】如圖 37B 所示，逆變換系統 3608 包括逆變換選擇器 3712 和一組逆變換邏輯單元 3714A-3714N。一組逆變換邏輯單元 3714A-3714N 可以包括硬體或軟體邏輯，以執行由變換系統 3604 的一組變換邏輯單元 3704A-3704N 進行的變換的逆變換。在一個實施例中，該組逆變換邏輯單元 3714A-3714N 被實施為可以由本文敘述的 GPGPU 執行的計算著色器或 CUDA 內核的形式的 GPGPU 可執行程式碼。

【0419】逆變換選擇器 3712 可以讀取應用於變換後的權重 3607 的變換類型 3605，並使用一組逆變換邏輯單元 3714A-3714N 中的適當的逆變換單元來執行逆變換，以產生權重矩陣 3609。根據與變換後的權重 3607 相關聯的變換類型 3605 來執行逆變換。在變換類型 3605 指示將不同的變

換應用於變換後的權重 3607 的不同子矩陣的情況下，將不同的逆變換應用於不同的部分。當將多個變換整體應用於子矩陣或權重矩陣時，將應用多個逆變換。

【0420】圖 38A-38B 繪示在圖形處理器上產生和使用變換後的矩陣的方法 3800、3820。圖 38A 繪示執行矩陣變換的方法 3800。圖 38B 繪示處理變換後矩陣的方法 3820。變換後的矩陣資料可以是例如由圖 36 和圖 37A 所示的變換系統 3604 輸出的變換後的權重資料。可以將變換後的矩陣資料編碼為稀疏編碼格式。在一實施例中，稀疏編碼格式可額外地使用資料壓縮邏輯來壓縮。在使用資料執行處理操作之前，圖 36 和圖 37B 所示的逆變換系統 3608 可以用於對矩陣資料進行逆變換。

【0421】如圖 38A 所示，方法 3800 包括用於計算裝置將矩陣資料載入到圖形處理器內的記憶體中 (3802)。矩陣資料可以被載入到主圖形記憶體中，矩陣資料的一部分可以被載入到圖形處理器的晶載記憶體中。晶載記憶體可以是圖形處理器內的變換邏輯 (諸如與變換系統 3604 相關聯的邏輯) 本地的。

【0422】在一實施例中，可以為矩陣資料指明特定的變換。例如，已知指明轉換對於矩陣資料是最佳的，可以旁路變換測試邏輯來應用該變換。如果為矩陣資料指明了特定的變換 (3803，是)，則計算裝置可以應用指明的變換 (3806)。可以指明一或多個變換。如果未指明變換 (3803，否)，則變換測試邏輯 3703 可用於判定最佳變換 (3804)。在

一個實施例中，用於矩陣資料的最佳變換是被判定為導致矩陣資料的最高稀疏度的變換。在一個實施例中，最佳變換是導致非零係數的最接近叢集的變換。在一個實施例中，最佳變換是當壓縮訓練資料時導致最高壓縮率的變換。然後可以經由數值變換邏輯將所判定的變換應用於矩陣資料(3806)。然後可以對變換後的矩陣資料進行編碼及/或壓縮(3808)。編碼和壓縮可以由計算裝置使用圖形處理器或經由在主機計算裝置上執行的軟體來執行。在使用圖形處理器執行編碼和壓縮的情況下，可以使用圖形處理器的硬體編碼器或編解碼器或經由在圖形處理器上執行的著色器或計算程式來執行編碼和壓縮。

【0423】如圖 38B 所示，方法 3820 包括用於計算裝置從記憶體讀取壓縮及/或編碼和變換的矩陣資料(3822)。記憶體可以是圖形處理器的記憶體、系統非揮發性儲存器或圖形處理器的非揮發性儲存器記憶體。可以讀取與矩陣資料相關聯的元資料，以指示已應用於變換後的矩陣資料的壓縮及/或編碼以及在壓縮之前已應用於矩陣資料的變換。使用壓縮及/或編碼元資料，計算裝置可以解壓縮及/或解碼變換後的矩陣資料(3824)。然後，計算裝置可以經由圖形處理器基於變換元資料將一或多個逆變換應用於變換後的矩陣資料(3826)。

【0424】計算裝置可以使用圖形處理器內的專用邏輯來對轉換後的矩陣資料進行解壓縮或解碼。計算裝置還可以使用在主機處理器或圖形處理器上執行的程式來解壓縮

或解碼變換後的矩陣資料。在一實施例中，在圖形處理器上執行的程式碼可以讀取儲存在記憶體中的壓縮及/或編碼和變換的資料，並且圖形處理器可以在提供資料給圖形處理器內的處理單元之前自動對資料進行解壓縮及/或解碼並對資料應用逆變換。然後，圖形處理器內的處理單元可以對矩陣資料執行計算操作以產生輸出資料(3828)。然後可以將處理單元產生的輸出資料寫入圖形處理器內的記憶體或系統記憶體。在一實施例中，如果輸出資料的期望特性(例如，壓縮率、稀疏特性)將被增加(3830)，則系統可以被組態以自動對輸出資料應用變換。

對壓縮資料執行脈衝式算術

【0425】 在一個實施例中，代替在使用脈動張量陣列2808處理資料之前解壓縮的資料，在一實施例中，張量管線接收壓縮或編碼格式的輸入以及敘述資料如何被壓縮或編碼的元資料。來自脈動張量陣列2808的輸出也可以被壓縮並且以壓縮格式被寫入記憶體。在一實施例中，壓縮和解壓縮還可包括應用本文敘述的變換和逆變換技術以增加資料可被壓縮到的壓縮率。

【0426】 圖39繪示被組態以對壓縮或編碼的資料進行操作的脈動張量陣列2808的版本。在一個實施例中，脈動張量陣列2808包括解碼器3312，其接收壓縮或編碼的資料以及相關聯的元資料3902。解碼器3312可以基於元資料對壓縮或編碼的資料進行解碼，以產生解壓縮的資料3508和

顯著圖 3506。旁路/閘邏輯 3510 可以基於由顯著圖 3506 指示的稀疏度及/或由顯著圖 3506 指明的特定零和非零元素判定旁路解壓縮的資料 3508 內的子矩陣及/或對脈動張量陣列 2808 內處理單元 3910 的某些處理元素進行時脈門控。處理單元 3910 輸出的資料可以由編碼器 3920 壓縮或編碼，並作為具有關聯的元資料 3928 的壓縮或編碼資料輸出。在一實施例中，編碼器 3920 可以包括圖 37A 的變換系統 3604 的數值變換邏輯，並且解碼器 3312 可以包括圖 37B 的逆變換系統 3608 的逆變換邏輯。在這樣的實施例中，元資料還將指示用於編碼資料的轉換類型。

【0427】圖 40 繪示對壓縮或編碼的資料執行脈衝式算術的方法 4000。可以藉由本文描述的硬體或軟體邏輯來執行方法 4000。在一個實施例中，硬體或軟體邏輯可以由在通用處理器上執行的軟體邏輯(例如驅動器邏輯)控制。

【0428】方法 4000 包括使計算系統執行包括從記憶體讀取壓縮及/或編碼資料的操作(方塊 4002)。然後，計算系統可以將壓縮及/或編碼的資料載入或饋送到張量加速器管線(方塊 4004)。張量加速器管線可為本文敘述之脈動管線，諸如脈動張量陣列或另一矩陣加速器架構。

【0429】然後，計算系統可以對壓縮及/或編碼的資料執行計算操作(方塊 4006)。可以基於由計算系統執行的程序碼指令來執行計算操作，例如像素著色器、頂點著色器、計算著色器、或 CUDA 程式。與計算系統內的圖形處理器的處理單元相關聯的解碼邏輯可以自動對壓縮及/或

編碼的資料進行解壓縮及/或解碼。例如，可以與由處理單元執行的載入操作一致地執行解壓縮及/或解碼。在一個實施例中，壓縮及/或編碼的資料的壓縮或編碼可以包括一或多個數值變換，其被應用來增加資料可以被壓縮到的壓縮率。在這樣的實施例中，解壓縮及/或解碼包括對資料應用數值逆變換。數值變換和數值逆變換是除了應用於資料的編碼演算法指明的任何變換或逆變換。

【0430】然後，計算系統可以壓縮及/或編碼張量加速器管線的輸出(方塊4008)。壓縮及/或編碼可以例如與由計算系統內的圖形處理器的處理單元執行的儲存操作一致地自動執行。在一個實施例中，根據圖36-38B中敘述的技術，可以在壓縮及/或編碼資料之前應用數值變換。然後，計算系統可以將壓縮及/或編碼的輸出寫入記憶體(方塊4010)。可以基於計算系統的微控制器內的韌體、或者基於計算系統內的硬體邏輯、經由儲存在非暫時性機器可讀取媒體中的指令來實現該方法的操作。

包括圖形處理器之額外的計算裝置

【0431】圖41係根據一實施例的包括圖形處理器4104之計算裝置4100的方塊圖。計算裝置4100可以是包括本文敘述的任何資料處理系統或子系統的計算裝置。計算裝置4100還可以被包括在通訊裝置內，或者被包括在諸如機頂盒(例如，基於網際網路的有線電視機頂盒等)、基於全球定位系統(GPS)的裝置等等的通訊裝置中。計算裝置4100

還可以是或被包括在諸如蜂巢式電話、智慧型手機、個人數位助理(PDA)、平板電腦、膝上型電腦、電子閱讀器、智慧電視、電視平台、可穿戴裝置(例如，眼鏡、手錶、手鐲、智慧卡、珠寶、衣服等)、媒體播放器等的行動計算裝置內。例如，在一實施例中，計算裝置4100包括採用積體電路(“IC”)(諸如系統單晶片(“SoC”或“SOC”))、在單一晶片上整合計算裝置4100的各種硬體及/或軟體組件的行動計算裝置。

【0432】計算裝置4100包括圖形處理器4104。圖形處理器4104代表本文敘述的任何圖形處理器。圖形處理器包括一或多個圖形引擎、圖形處理器核心以及本文所述的其他圖形執行資源。此類圖形執行資源可以以下形式呈現：包括但不限於執行單元、著色器引擎、片段處理器、頂點處理器、串流多處理器、圖形處理器叢集、或適用於圖形資源或圖像資源處理或在異構處理器中執行通用計算操作的任何計算資源集合。

【0433】在一實施例中，圖形處理器4104包括快取4114，其可為單一快取，或分為快取記憶體的多個片段，包括但不限於任意數量的L1、L2、L3或L4快取、渲染快取、深度快取、取樣器快取及/或著色器單元快取。在一些實施例中，圖形處理器4104包括GPGPU引擎4144，其包括工作負載處置器4140、張量加速器4134及變換/逆變換邏輯4124。工作負載處置器4140可以排程工作負載操作以在張量加速器4134和GPGPU引擎4144上執行。工作負載單

元 4124 可以具有硬體邏輯單元，包括但不限於圖 27 的排程器控制器 2722。在一個實施例中，張量加速器 4134 包括或者是圖 27 中的張量加速器 2723 的版本。在一實施例中，張量加速器 4134 可包括本文所述之脈動張量陣列（例如，脈動張量陣列 2808）。

【0434】如圖所示，在一實施例中，除了圖形處理器 4104 之外，計算裝置 4100 還可包括任何數量和類型的硬體組件及 / 或軟體組件，包括但不限於應用程式處理器 4106、記憶體 4108、輸入 / 輸出 (I/O) 源 4110。應用程式處理器 4106 可以與硬體圖形管線交互以共享圖形管線功能。經處理的資料儲存於在硬體圖形管線中的緩衝器，以及狀態資訊儲存在記憶體 4108 中。可以將所得資料傳送到顯示控制器，以經由諸如圖 16B 的顯示裝置 1618 的顯示裝置輸出。顯示裝置可為諸如陰極射線管 (CRT)、薄膜電晶體 (TFT)、液晶顯示器 (LCD)、有機發光二極體 (OLED) 陣列等各種類型，並且可以組態以經由圖形使用者介面向使用者顯示資訊。

【0435】應用程式處理器 4106 可以包括一或多個處理器，諸如圖 1 的處理器 102，並且可以是中央處理單元 (CPU)，其可至少部分地用於執行計算裝置 4100 的作業系統 (OS) 4102。OS 4102 可以用作計算裝置 4100 的硬體及 / 或實體資源與一或多個使用者之間的介面。OS 4102 可以包括用於計算裝置 4100 中的各種硬體裝置的驅動器邏輯，包括圖形驅動器邏輯 4122，諸如圖 23 的使用者模式圖形驅動

器 2326 及 / 或內核模式圖形驅動器 2329。

【0436】可以想到，在一些實施例中，圖形處理器 4104 可以作為應用程式處理器 4106 的一部分 (諸如實體 CPU 封裝的一部分) 存在，在這種情況下，記憶體 4108 的至少一部分可以由應用程式處理器 4106 和圖形處理器 4104 共享，儘管記憶體 4108 的至少一部分可以是圖形處理器 4104 所獨占的，或者圖形處理器 4104 可以具有單獨的記憶體儲存。記憶體 4108 可包含有緩衝器 (例如，訊框緩衝器) 之預分配區；然而，所屬技術領域中具有通常知識者應瞭解的是實施例並不如此受限，並且可使用可存取下圖形管線之任何記憶體。記憶體 4108 可以包括各種形式的隨機存取記憶體 (RAM) (例如，SDRAM，SRAM 等)，其包括利用圖形處理器 4104 來渲染桌面或 3D 圖形場景的應用。

【0437】記憶體控制器集線器 (諸如，圖 14 之記憶體控制器 1416) 可以存取記憶體 4108 中的資料並將其轉發至圖形處理器 4104 以進行圖形管線處理。記憶體 4108 可用於計算裝置 4100 內的其他組件。例如，從計算裝置 4100 之各種 I/O 源 4110 接收的任何資料 (例如，輸入圖形資料) 可在軟體程式或應用程式的實施方式中，在它們被一或多個處理器 (例如，應用程式處理器 4106) 操作之前，暫時將它們排入記憶體 4108。類似的是，軟體程式判定之資料應該透過此等計算系統介面之一者自計算裝置 4100 發送至外側實體，或儲存到內部儲存元件裡，在傳送或儲存之前，通常先於記憶體 4108 中暫時排入佇列。

【0438】I/O源可以包括諸如觸控螢幕、觸控面板、觸控板、虛擬或常規的鍵盤、虛擬或常規的滑鼠、埠、連接器、網路裝置等的裝置，並且可以圖14所示經由平台控制器集線器1430來附接。額外地，I/O源4110可以包括一或多個I/O裝置，其被實現用於向及/或從計算裝置4100(例如，網路配接器)傳輸資料；或者，用於計算裝置4100內的大規模非揮發性儲存器(例如，硬碟驅動器)。包括字母數字和其他按鍵的使用者輸入裝置可被用來傳遞資訊和命令選擇至圖形處理器4104。另一類型的使用者輸入裝置為游標控制，諸如滑鼠、軌跡球、觸控螢幕、觸控板、或游標方向鍵，其用來傳遞方向資訊與命令選擇至GPU，並且用來控制顯示裝置上的游標移動。計算裝置4100的相機及麥克風陣列可運用於觀測手勢、記錄音訊與視訊，以及運用於接收及傳送視覺與音訊命令。

【0439】組態為網路介面的I/O源4110可以提供對網路的存取，諸如LAN、廣域網路(WAN)、都會區網路(WAN)、個人區域網(PAN)、藍牙、雲端網路、蜂巢或行動網路(例如，第三代(3G)、第四代(4G)等)、企業內部網路、網際網路等。網路介面可包括(例如)具有一或多個天線的無線網路介面。網路介面可包括例如經由網路纜線與遠端裝置通訊之有線網路介面，網路纜線可以例如是乙太網路纜線、同軸纜線、光纖纜線、串列纜線或平行纜線。

【0440】網路介面可藉由例如符合IEEE 802.11標準提供對LAN之存取，及/或無線網路介面可藉由例如符合

藍芽標準提供對個人區域網路之存取。也可支持其他包括先前及後續版本之標準的無線網路介面及/或協定。除了經由無線 LAN 標準進行通訊以外、或對其取而代之外，網路介面可使用例如分時多重存取 (TDMA) 協定、全球行動通訊系統 (GSM) 協定、分碼多重存取 (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。

【0441】要了解就某些實施方式，比上述實例更少或更多配備之系統可能較佳。因此，計算裝置 4100 之組態可隨實施方式之不同而變，端視諸如價格限制條件、效能要求、技術改良、或其他情況等許多因素而定。實例可包括 (非限制於) 行動裝置、個人數位助理、行動計算裝置、智慧型手機、蜂巢式電話、手持話機、單向呼叫器、雙向呼叫器、傳訊裝置、電腦、個人電腦 (PC)、桌上型電腦、膝上型電腦、筆記型電腦、手持電腦、平板電腦、伺服器、伺服器陣列或伺服器場、網頁伺服器、網路伺服器、網際網路伺服器、工作站、迷你電腦、主機架電腦、超級電腦、網路設備、網路家電、分散式計算系統、多處理器系統、處理器為基之系統、消費性電子、可編程消費性電子、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、電橋、切換、機器或以上的組合。

【0442】實施例可實施成下列任何一者或其組合：使用主機板互連之一或多個微晶片或積體電路、固線式邏輯、由記憶體裝置所儲存並由微處理器所執行之軟體、韌

體、特定應用積體電路(ASIC)及/或現場可程式閘陣列(FPGA)。「邏輯」一詞可包括例如軟體或硬體及/或軟體與硬體之組合。

【0443】實施例可被提供作為例如電腦程式產品，可包括其上儲存有機器可執行指令之一或多個機器可讀媒體，此等機器可執行指令在由諸如電腦、電腦網路或其他電子裝置等一或多個機器執行時，可導致一或多個機器根據本文中所述實施例進行操作。機器可讀取媒體可包括但不限制於軟式磁片、光碟、CD-ROM(光碟唯讀記憶體)、以及磁光碟、ROM、RAM、EPROM(可抹除可編程唯讀記憶體)、EEPROM(電氣可抹除可編程唯讀記憶體)、磁性或光學卡、快閃記憶體或適用於儲存機器可執行指令之其他類型的非暫態機器可讀取媒體。

【0444】此外，可下載實施例作為電腦程式產品，其中可經由一通訊鏈路(例如，數據機及/或網路連線)，藉由具體實施成載波或其他傳播媒體及/或由其所調變之一或多個資料訊號，將程式從遠端電腦(例如，伺服器)轉移至提出請求之電腦(例如，用戶端)。

【0445】本文敘述的實施例提供了一種指令和相關聯的邏輯，以使GPGPU程式碼能夠存取專用硬體邏輯以加速內積操作。以下條款及/或示例屬於特定實施例或其實例。在一或多個實施例中範例中的特定可在任何地方使用。不同實施例或實例的各種特徵可以與所包括的一些特徵以及排除在外的一些特徵進行各種組合，以適於各種不

同的應用。實例可以包括諸如方法的請求標的，用於執行該方法的動作的手段，至少一機器可讀取媒體，該至少一機器可讀取媒體包括指令，根據本文描述的實施例和實例當機器執行該指令時，該指令導致該機器執行該方法、設備或系統的動作。各種組件可以是用於執行所敘述的操作或功能的裝置。

【0446】一個實施例提供了一種包含一或多個包括圖形處理器的處理器的計算裝置。圖形處理器包括處理資源，處理資源包括張量加速器。張量加速器係組態以執行數值操作以訓練類神經網路模型並產生與該類神經網路模型相關聯之第一權重矩陣。圖形處理器額外地包括權重變換器，以將數值變換應用於第一權重矩陣以產生一組變換後的權重和變換類型。變換類型標識應用於第一權重矩陣的數值變換。第一權重矩陣是稀疏矩陣，並且變換後的權重壓縮到比第一權重矩陣更高的壓縮比。

【0447】在各種實施例中，權重轉換器是圖形處理器的固定功能硬體單元，或者經由圖形處理器的可編程硬體來實現。權重變換器包括變換選擇器和變換測試器。變換測試器可組態以將第一數值變換應用於第一權重矩陣的至少一部分以產生第一測試變換資料、將第二數值變換應用於第一權重矩陣的至少一部分以產生第二測試變換資料、基於對第一測試變換資料和第二測試變換資料的分析來判定可壓縮性度量，並將推薦的變換發送到變換選擇器。可以測試的數值變換的具體數量沒有限制。測試轉換可以並

行執行。然後，變換選擇器可以基於推薦的變換選擇數值變換以應用於第一權重矩陣。除了推薦的變換之外，還可以選擇數值變換。可以從包括離散餘弦變換、離散正弦變換、位元翻轉變換及位元旋轉變換的一組數值變換中選擇數值變換。在另一個實施例中，圖形處理器還包括逆權重變換器，以將數值逆變換應用於變換後的權重，以產生第二權重矩陣。藉由經由與變換權重組相關聯的變換類型變換的逆權重的逆變換選擇器來識別要執行的數值逆變換。

【0448】一個實施例提供了一種方法，其包含在包括張量加速器的圖形處理器上，將矩陣資料以及與該矩陣資料相關聯的元資料載入至該張量加速器，其中該元資料指示施加至該矩陣資料的數值變換，由該張量加速器對該矩陣資料執行數值逆變換，由與該矩陣資料相關聯的該元資料指示該數值逆變換，在執行該數值逆變換之後，執行在張量加速器內的一或多個計算操作，以及將一或多計算操作的輸出寫入至該圖形處理器的記憶體。

【0449】在進一步實施例中，該方法進一步包含再對該矩陣資料執行該數值逆變換之前，對該張量加速器內的該矩陣資料解壓縮或解碼，以及在對矩陣資料執行該數值變換之後，對該張量加速器內的該輸出壓縮或編碼。在一實施例中，將該一或多個計算操作的輸出寫入至該圖形處理器之該記憶體包括將該輸出寫入至該圖形處理器之快取記憶體及/或將該輸出寫入至該圖形處理器之主記憶體。

【0450】在將該輸出寫入至該圖形處理器之該記憶體

之前，亦可將數值變換應用於該一或多個計算操作的該輸出。該數值變換可以指明數值變換。應用該數值變換亦可包括由該張量加速器產生用於多個數值變換的變換測試度量，基於該變換測試度量選擇數值變換，以及將所選擇的數值變換應用於該輸出。該選擇的數值變換可以從包括離散餘弦變換、離散正弦變換、位元翻轉變換及位元旋轉變換的一組數值變換中選擇。

【0451】一個實施例提供圖形處理器，其包含：處理資源，其包括張量加速器，其中該張量加速器係用以執行數值操作以訓練類神經網路模型並產生與該類神經網路模型相關聯之第一權重矩陣，權重變換器，其將數值變換應用於該第一權重矩陣以產生一組變換的權重和變換類型，其中該變換類型識別應用至該第一權重矩陣的該數值變換，其中該第一權重矩陣係稀疏矩陣並且該變換的權重壓縮到比該第一權重矩陣更高的壓縮比，以及逆權重變換器，其將數值逆變換應用於該變換的權重，以產生第二權重矩陣，其中要執行的該數值逆變換係經由與該組變換的權重相關聯的該變換類型來識別的。

【0452】在進一步實施例中，該權重變換器包括變換選擇器和變換測試器，該變換測試器用以將第一數值變換應用於該第一權重矩陣的至少一部分以產生第一測試變換資料、將第二數值變換應用於該第一權重矩陣的至少一部分以產生第二測試變換資料、基於對該第一測試變換資料和該第二測試變換資料的分析來判定可壓縮性度量，並將

推薦的變換發送到該變換選擇器，該變換變換器用以基於該推薦的變換選擇應用至該第一權重矩陣的數值變換。該數值變換係從包括離散餘弦變換、離散正弦變換、位元翻轉變換及位元旋轉變換的一組數值變換中選擇。

【0453】前面的敘述和圖式應被認為是說明性的而不是限制性的。本領域的技術人員將理解，在不脫離如申請專利範圍中敘述之特徵的更寬泛的精神和範圍的情況下，可以對本文所述的實施例進行各種修改和改變。

【符號說明】

【0454】

100:計算系統
101:處理子系統
102:處理器
104:系統記憶體
105:記憶體集線器
106:通訊鏈路
107:I/O集線器
108:輸入裝置
110A,110B:顯示裝置
111:I/O子系統
112:並行處理器
113:通訊鏈路
114:系統儲存單元

116:I/O開關

118:網路配接器

119:無線網路配接器

120:外接裝置

200:並行處理器

202:並行處理單元

204:I/O單元

206:主介面

208:前端

210:排程器

212:處理叢集陣列

214A~214N:叢集

216:記憶體交叉開關

218:記憶體介面

220A~220N:分割單元

221:L2快取

222:並行處理器記憶體

224A~224N:記憶體單元

225:框緩衝器介面

226:ROP

227:CODEC

232:管線管理器

234:圖形多處理器

236:紋理單元

240:資料交叉開關

242:preROP

245:MMU

248:第一階 1(L1)快取

252:指令快取

254:指令單元

256:位址映射單元

258:暫存器檔案

262:GPGPU核心

263:射線追蹤核心

266:載入/儲存單元

268:記憶體及快取互連

270:共享記憶體

272:快取記憶體

325:圖形多處理器

327:互連結構

330:共同指令快取

332A,332B:指令單元

334A,334B:暫存器檔案

336A,336B:GPGPU核心

337A,337B:張量核心

338A,338B:射線追蹤核心

340A,340B:載入/儲存單元

342,358A,358B:快取記憶體

344A,344B,360A~360D:紋理單元

346,353:共享記憶體

350:圖形多處理器

352:互連結構

354:指令快取

356A,356B,356B,356D:執行資源

361:CPU

362:I/O裝置

363:輸入/輸出(I/O)電路

364:I/O記憶體管理單元(IOMMU)

365A~365N:多核心群組

366:系統記憶體

367:記憶體控制器

368:排程器/調度器

369:暫存器檔案

370:圖形核心

371:張量核心

372:射線追蹤核心

373:L1快取及共享記憶體單元

374:紋理單元

375:L2快取

380:GPU

401~402:處理器記憶體

405~406:多核心處理器

407:處理器

410~413:GPU

420~423:GPU記憶體

425:代理電路

430A,430B:記憶體互連

431~432:圖形處理引擎

433~434:圖形記憶體

435,437:介面

436:加速器積體電路

438,462A~462D,456:快取

439:MMU

440A~440D,442A~442B,443:高速鏈路

441:系統記憶體

445:暫存器

446:圖形加速模組

447:中斷管理電路

448:背景管理電路

450A~450D:GPU記憶體互連

460A~460D:核心

461A~461D:轉換後備緩衝器

464:同調匯流排

480:應用程式

481:GPU調用

482:應用程式有效位址空間

483:程序元件

484:工作描述符(WD)

485:虛擬位址空間

486:分段/頁面表

490:加速器整合片

491:WD提取單元

493:有效位址

494A~494E:偏移/同調管理電路

495:作業系統

496:超管理器

497:共享的I/O

498:超管理器真實位址空間

499:程序元件列表

500:圖形處理管線

502:資料組譯器

504:頂點處理單元

506,514,518:基元組譯器

508:鑲嵌控制處理單元

510:鑲嵌單元

512:鑲嵌評估處理單元

516:幾何處理單元

520:視埠縮放、剔除及修剪單元

522:柵格化器

526:柵格操作單元

528:記憶體介面
600:機器學習軟體堆疊
602:機器學習應用程式
604:機器學習框架
606:計算框架
608:GPGPU驅動器
610:GPGPU硬體
700:通用圖形處理單元
702:主介面
704:全域排程器
706A~706H:處理叢集
708:快取記憶體
709:I/O集線器
710:GPU鏈路
712A~712B:記憶體控制器
714A~714B:記憶體
800:多GPU計算系統
802:處理器
804:主介面開關
806A~806D:GPGPU
816:P2P GPU鏈路
902:輸入
904,906,912,914:卷積層
908:完全連接層

916:卷積階段
918:檢測器階段
920:池化階段
922:下一層
1000:例示性遞迴類神經網路
1002:輸入層
1004:隱藏層
1005:回饋機制
1006:輸出層
1102:訓練資料集
1104:訓練框架
1106:未訓練類神經網路
1108:經訓練類神經網路
1112:新資料
1114:結果
1202:模型並行性
1204:資料並行性
1206:模型和資料並行性之組合
1220,1300:系統單晶片 (SoC)
1222:處理器核心
1230:管理介面
1240:記憶體裝置
1245:GPU
1250:儲存裝置

1260A~1260B:網路埠

1270:主介面

1275:I/O介面

1302:媒體處理器

1304:視覺處理器

1305:晶載記憶體

1306:GPGPU

1308:多核心處理器

1400:處理系統

1402:處理器

1404:快取記憶體

1406:暫存器檔案

1407:處理器核心

1408:圖形處理器

1409:指令集

1410:介面匯流排

1411:顯示裝置

1412:加速器

1416:積體記憶體控制器

1418:外部圖形處理器

1419:外部加速器

1420:記憶體裝置

1421:指令資料

1422:資料

1424:資料儲存裝置
1425:觸控感測器
1426:無線收發器
1428:韌體介面
1430:平台控制器集線器(PCH)
1434:網路控制器
1440:舊式I/O控制器
1442:通用串列匯流排(USB)控制器
1443:鍵盤和滑鼠
1444:相機
1446:音頻控制器
1500:處理器
1502A~1502N:核心
1504A~1504N:快取單元
1506:共享快取單元
1508:圖形處理器
1510:系統代理核心
1511:顯示控制器
1512:環狀互連單元
1513:I/O鏈路
1514:記憶體控制器
1516:匯流排控制器單元
1518:嵌入式記憶體模組
1519:圖形處理器核心

1521A~1521F:多個子核心

1522A~1522F,1524A~1524F:執行單元(EU)陣列

1523A~1523F:執行緒分派和執行緒間通訊(TD/IC)邏輯

1525A~1525F:3D取樣器

1526A~1526F:媒體取樣器

1527A~1527F:著色器處理器

1528A~1528F:共享本地記憶體(SLM)

1530:固定功能塊

1531:幾何/固定功能管線

1532:圖形SoC介面

1533:圖形微控制器

1534:媒體管線

1535:共享功能邏輯

1536:共享及/或快取記憶體

1537:幾何/固定功能管線

1538:額外的固定功能邏輯

1546:CPU

1553:L2快取

1554:L1快取

1555:指令快取

1556:共享記憶體

1557:命令處理器

1558:執行緒分派器

1560A~1560N:計算單元

- 1561:向量暫存器
- 1562:純量暫存器
- 1563:向量邏輯單元
- 1564:純量邏輯單元
- 1565:本地共享記憶體
- 1566:程式計數器
- 1567:恆定快取
- 1568:記憶體控制器
- 1569:內部直接記憶體存(DMA)控制器
- 1570:GPGPU
- 1571,1572:記憶體
- 1600:圖形處理器
- 1602:顯示控制器
- 1603:區塊影像轉移(BLIT)引擎
- 1604:複製引擎
- 1606:視訊編碼解碼引擎
- 1610:圖形處理引擎
- 1610A~1610D:圖形引擎磚片
- 1612:3D管線
- 1614:記憶體介面
- 1615:3D/媒體子系統
- 1616:媒體管線
- 1618:顯示裝置
- 1620,1630:圖形處理器

1622,1632:圖形處理引擎叢集

1623A~1623F:磚片互連

1624:結構互連

1625A~1625D:記憶體互連

1626A~1626D:記憶體模組或記憶體裝置

1628:主介面

1636:L3快取

1640A~1640D:計算引擎磚片

1642:網路介面

1644:實體層互連

1703:命令串流器

1710:圖形處理引擎

1714:圖形核心陣列

1715A,1715B:圖形核心

1716,1720:共享功能邏輯

1718:統一返回緩衝器

1721:取樣器邏輯

1722:數學邏輯

1723:執行緒間通訊(ITC)

1725:快取

1800:執行邏輯

1802:著色器處理器

1804:執行緒分派器

1805:射線追蹤器

1806:指令快取

1807A~1807N:執行緒控制邏輯

1808A~1808N:圖形執行單元

1809A~1809N:融合圖形執行單元

1810:取樣器

1811:共享本地記憶體

1812:資料快取

1814:資料埠

1822:執行緒仲裁器

1824:通用暫存器檔案陣列 (GRF)

1826:架構暫存器檔案陣列 (ARF)

1830:發送單元

1832:分支單元

1834:浮點單元 (FPU)

1835:SIMD ALU

1837:指令提取單元

1900:執行單元

1901:執行緒控制單元

1902:執行緒狀態單元

1903:指令提取/預提取單元

1904:指令解碼單元

1906:暫存器檔案

1907:發送單元

1908:分支單元

1910:計算單元

1911:ALU

1912:脈動陣列

1913:數學單元

2000:圖形處理器指令格式

2010:128位元指令格式

2012:指令運算碼

2013:指標欄位

2014:指令控制欄位

2016:執行大小欄位

2018:目的地

2020:src0

2022:src1

2024:SRC2

2026:存取/位址模式欄位

2030:64位元緊湊指令格式

2040:運算碼解碼

2042:移動和邏輯運算碼群組

2044:流程控制指令群組

2046:雜項指令群組

2048:並行數學指令群組

2050:向量數學群組

2100:圖形處理器

2102:環狀互連

2103:命令串流器
2105:頂點提取器
2107:頂點著色器
2111:外殼著色器
2113:鑲嵌器
2117:領域著色器
2119:幾何著色器
2120:幾何管線
2123:串流輸出單元
2129:截波器
2130:媒體管線
2131:執行緒分配器
2134:視訊前端
2137:媒體引擎
2140:顯示引擎
2141:2D引擎
2143:顯示控制器
2150:執行緒執行邏輯
2151,2158:快取
2152A~2152B:執行單元
2154:取樣器
2156:資料埠
2170:渲染輸出管線
2173:光柵化器及深度測試組件

2175:共享 L3 快取
2177:像素操作組件
2178:渲染快取
2179:深度快取
2200:圖形處理器命令格式
2202:客戶端
2204:命令運算碼(opcode)
2205:子運算碼
2206:資料
2208:命令大小
2210:圖形處理器命令序列
2212:管線清除命令
2213:管線選擇命令
2214:管線控制命令
2216:返回緩衝器狀態
2220:管線判定
2222:3D管線
2224:媒體管線
2230:3D管線狀態
2232:3D基元
2234:執行
2240:媒體管線狀態
2242:媒體物件命令
2244:執行命令

2300:資料處理系統

2310:3D圖形應用程式

2312:著色器指令

2314:可執行指令

2316:圖形物件

2320:作業系統

2322:圖形 API

2324:前端著色器編譯器

2326:使用者模式圖形驅動器

2327:後端著色器編譯器

2328:作業系統內核模式功能

2329:內核模式圖形驅動器

2330:處理器

2332:圖形處理器

2334:通用處理器核心

2350:記憶體

2400:IP核心開發系統

2410:軟體模擬

2412:模擬模型

2415:暫存器轉移層級(RTL)設計

2420:硬體模型

2430:設計設施

2440:非揮發性記憶體

2450:有線連接

2460:無線連接
2465:製造設施
2470:積體電路封裝總成
2472,2474:硬體邏輯
2473:互連結構
2475:記憶體小晶片
2480:基材
2482:電橋
2483:封裝互連
2485:結構
2487:電橋結構
2489:有源中介層
2490:封裝總成
2491:I/O
2492:快取記憶體
2493:硬體邏輯
2494:封裝總成
2495:可互換小晶片
2496,2498:基底小晶片
2497:電橋互連
2500:積體電路
2505:應用程式處理器
2510:圖形處理器
2515:影像處理器

2520:視訊處理器

2525:USB控制器

2530:UART控制器

2535:SPI/SDIO控制器

2540:I2S/I2C控制器

2545:顯示裝置

2550:高解析度多媒體介面(HDMI)控制器

2555:行動產業處理器介面(MIPI)顯示介面

2560:快閃記憶體子系統

2565:記憶體控制器

2570:嵌入式安全性引擎

2605:頂點處理器

2610,2640:圖形處理器

2615A~2615N:片段處理器

2620A,2620B:記憶體管理單元(MMU)

2625A,2625B:快取

2630A,2630B:電路互連

2645:核心間任務管理器

2655A~2655N:著色器核心

2658:鋪磚單元

2700:資料處理系統

2702:處理器

2710:統一記憶體

2712:系統記憶體

2714A:來源碼
2714B:已編譯碼
2715:編譯器
2716:運行時庫
2718:GPGPU記憶體
2720:GPGPU
2721:提取和解碼單元
2722:排程器控制器
2723:張量加速器
2724A~2724N:計算區塊
2725:暫存器
2726:功率和效能模組
2727:快取
2730:L3資料快取
2732:共享本地記憶體
2800:指令管線
2802:接收欄位磚片
2804:輸入量緩衝器
2805:點矩陣操作
2806:輸出緩衝器
2808:脈動張量陣列
2830:記憶體
2900,2901,2902:來源
2904,2906:輸入資料緩衝器

2912:功能單元

2922:累加器

2923A~2329D:乘法器

2924:加法器

2926:輸入值

2928:額外 SIMD 通道

3000:處理資源

3006:暫存器檔案

3010:計算

3011:ALU

3012:張量加速器

3102,3104:矩陣

3105:單一非零值子矩陣

3112,3114:子矩陣圖

3120:記憶體

3121:負載 B 濾波器

3122:矩陣 B 載入單元

3124:矩陣 B 饋送單元

3126:矩陣 A 載入單元

3127:負載 A 濾波器

3128:矩陣 A 饋送單元

3130:功能單元

3202:子矩陣旁路訊息

3204:旁路的子矩陣資料

3300:計算架構
3302:計算區塊
3304:高速暫存緩衝器
3306:DMA控制器
3308:記憶體
3312:解碼器
3314:壓縮資料
3316:編碼器
3410:輸出資料緩衝器
3420:編碼佈局
3422:標頭
3424:顯著圖
3426:非零值
3430:編碼的位元串流
3506:顯著圖
3508:解壓縮的資料
3510:旁路/閘邏輯
3600:系統
3602:訓練系統
3603,3609:權重矩陣
3604:變換系統
3605:變換類型
3607:變換後的權重
3608:逆變換系統

3610:推理系統

3701,3711:記憶體

3702:變換選擇器

3703:變換測試邏輯

3704A~3704N:變換邏輯單元

3712:逆變換選擇器

3714A~3714N:逆變換邏輯單元

3800,3820,4000:方法

3902,3928:壓縮或編碼的資料以及相關聯的元資料

3910:處理單元

3920:編碼器

4002,4004,4006,4008,4010:方塊

4100:計算裝置

4102:作業系統(OS)

4104:圖形處理器

4106:應用程式處理器

4108:記憶體

4110:輸入/輸出(I/O)源

4114:快取

4122:圖形驅動器邏輯

4124:變換/逆變換邏輯

4134:張量加速器

4140:工作負載處置器

4144:GPGPU引擎

【發明申請專利範圍】

【請求項 1】一種用於稀疏資料之脈動式算術的計算裝置，其包括：

一或多個處理器，其包括圖形處理器，該圖形處理器包括：

處理資源，其包括張量加速器，其中該張量加速器係用以執行數值操作以訓練類神經網路模型並產生與該類神經網路模型相關聯之第一權重矩陣；以及

權重變換器，其將數值變換應用於該第一權重矩陣以產生一組變換的權重和變換類型，其中該變換類型識別應用至該第一權重矩陣的該數值變換，其中該第一權重矩陣係稀疏矩陣並且該變換的權重壓縮到比該第一權重矩陣更高的壓縮比。

【請求項 2】如請求項 1 之計算裝置，其中該權重變換器係該圖形處理器之固定功能硬體單元。

【請求項 3】如請求項 1 之計算裝置，其中該權重變換器係經由該圖形處理器之可編程硬體實施。

【請求項 4】如請求項 1 之計算裝置，其中該權重變換器包括變換選擇器和變換測試器，該變換測試器用以：

將第一數值變換應用於該第一權重矩陣的至少一部分以產生第一測試變換資料；

將第二數值變換應用於該第一權重矩陣的至少一部分以產生第二測試變換資料；

基於對該第一測試變換資料和該第二測試變換資料的

分析來判定可壓縮性度量；以及

將推薦的變換發送到該變換選擇器。

【請求項 5】如請求項 4 之計算裝置，該變換選擇器進一步用以基於該推薦的變換選擇數值變換以應用於該第一權重矩陣。

【請求項 6】如請求項 5 之計算裝置，其中該數值變換係從包括離散餘弦變換、離散正弦變換、位元翻轉變換及位元旋轉變換的一組數值變換中選擇。

【請求項 7】如請求項 6 之計算裝置，該圖形處理器額外地包括：

逆權重變換器，其將數值逆變換應用於該組變換的權重，以產生第二權重矩陣，其中要執行的該數值逆變換係經由與該組變換的權重相關聯的該變換類型來識別的。

【請求項 8】如請求項 1-7 中任一項之計算裝置，其中該張量加速器包括處理元件的脈動陣列。

【請求項 9】一種用於稀疏資料之脈動式算術的方法，包括：

在包括張量加速器的圖形處理器上：

將矩陣資料和與該矩陣資料相關聯的元資料載入至該張量加速器，其中該元資料指示應用至該矩陣資料的數值變換；

由該張量加速器對該矩陣資料執行數值逆變換，該數值逆變換由與該矩陣資料相關聯的該元資料指示；

在執行該數值逆變換之後，執行該張量加速器內

的一或多個計算操作；以及

將該一或多個計算操作之輸出寫入至該圖形處理器的記憶體。

【請求項 10】如請求項 9 之方法，額外地包括在對該矩陣資料執行該數值逆變換之前將該張量加速器內的該矩陣資料進行解壓縮或解碼。

【請求項 11】如請求項 10 之方法，額外地包括在對該矩陣資料執行該數值變換之後將該張量加速器內的該矩陣資料進行壓縮或編碼。

【請求項 12】如請求項 9 之方法，其中將該一或多個計算操作之輸出寫入至該圖形處理器之該記憶體包括將該輸出寫入至該圖形處理器之快取記憶體。

【請求項 13】如請求項 9 之方法，其中將該一或多個計算操作之輸出寫入至該圖形處理器之該記憶體包括將該輸出寫入至該圖形處理器之主記憶體。

【請求項 14】如請求項 9 之方法，額外地包括在將該輸出寫入至該圖形處理器之該記憶體之前，將數值變換應用至該一或多個計算操作的該輸出。

【請求項 15】如請求項 14 之方法，其中應用該數值變換至該一或多個計算操作的該輸出包含將指明的數值變換應用至該輸出。

【請求項 16】如請求項 14 之方法，其中應用該數值變換至該一或多個計算操作之該輸出包含由該張量加速器產生用於多個數值變換的變換測試度量、基於該變換測試度

量選擇數值變換、以及將該選擇的數值變換應用於該輸出。

【請求項 17】如請求項 16 之方法，其中該選擇的數值變換係從包括離散餘弦變換、離散正弦變換、位元翻轉變換及位元旋轉變換的一組數值變換中選擇。

【請求項 18】一種用於稀疏資料之脈動式算術的圖形處理系統，其包括用以執行如請求項 9-17 中任一項之方法的手段。

【請求項 19】一種用於稀疏資料之脈動式算術的非暫時性機器可讀取媒體，該非暫時性機器可讀取媒體儲存資料，當由一或多個機器讀取時，使該一或多個機器製造一或多個積體電路以執行請求項 9-17 中任一項所述方法的操作。

【請求項 20】一種用於稀疏資料之脈動式算術的圖形處理器，包含：

處理資源，其包括張量加速器，其中該張量加速器係用以執行數值操作以訓練類神經網路模型並產生與該類神經網路模型相關聯之第一權重矩陣；

權重變換器，其將數值變換應用於該第一權重矩陣以產生一組變換的權重和變換類型，其中該變換類型識別應用至該第一權重矩陣的該數值變換，其中該第一權重矩陣係稀疏矩陣並且該變換的權重壓縮到比該第一權重矩陣更高的壓縮比；以及

逆權重變換器，其將數值逆變換應用於該組變換的權

重，以產生第二權重矩陣，其中要執行的該數值逆變換係經由與該組變換的權重相關聯的該變換類型來識別的。

【請求項 21】如請求項 20 之圖形處理器，其中該權重變換器包括變換選擇器和變換測試器，該變換測試器用以：

將第一數值變換應用於該第一權重矩陣的至少一部分以產生第一測試變換資料；

將第二數值變換應用於該第一權重矩陣的至少一部分以產生第二測試變換資料；

基於對該第一測試變換資料和該第二測試變換資料的分析來判定可壓縮性度量；以及

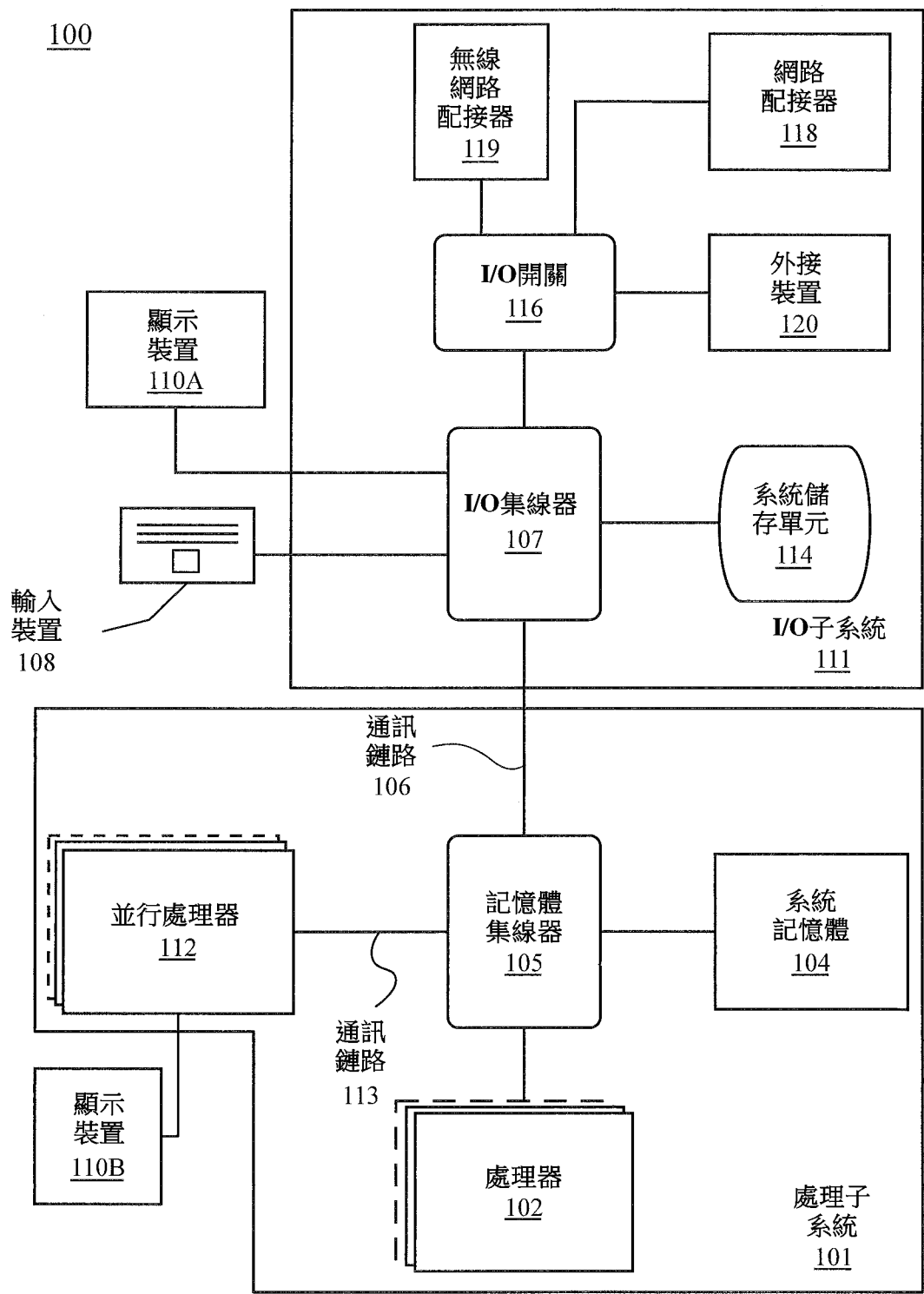
將推薦的變換發送到該變換選擇器。

【請求項 22】如請求項 21 之圖形處理器，該變換選擇器進一步用以基於該推薦的變換選擇數值變換以應用於該第一權重矩陣。

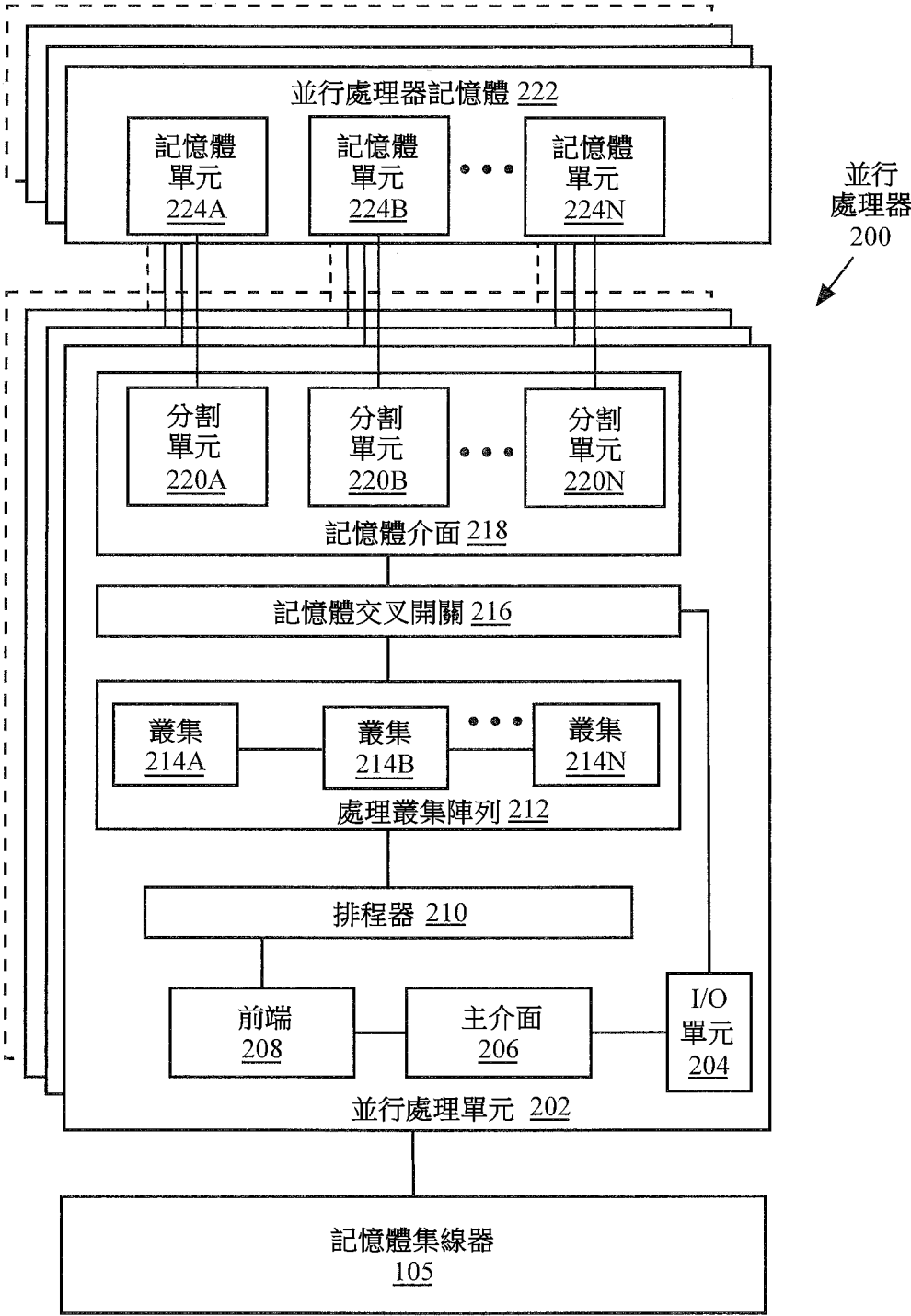
【請求項 23】如請求項 22 之圖形處理器，其中該數值變換係從包括離散餘弦變換、離散正弦變換、位元翻轉變換及位元旋轉變換的一組數值變換中選擇。

【請求項 24】如請求項 20-23 中任一項之圖形處理器，其中該張量加速器包括處理元件的脈動陣列。

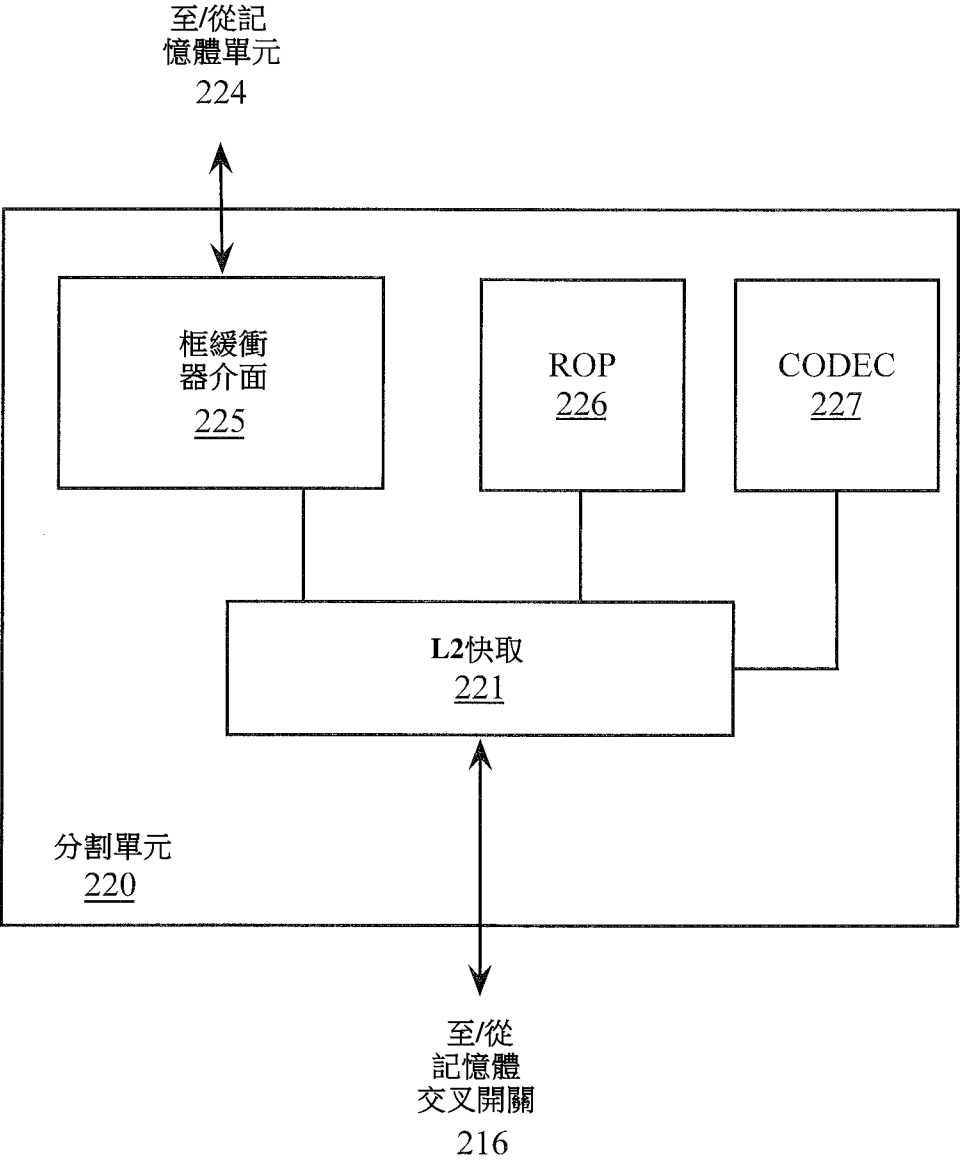
【發明圖式】



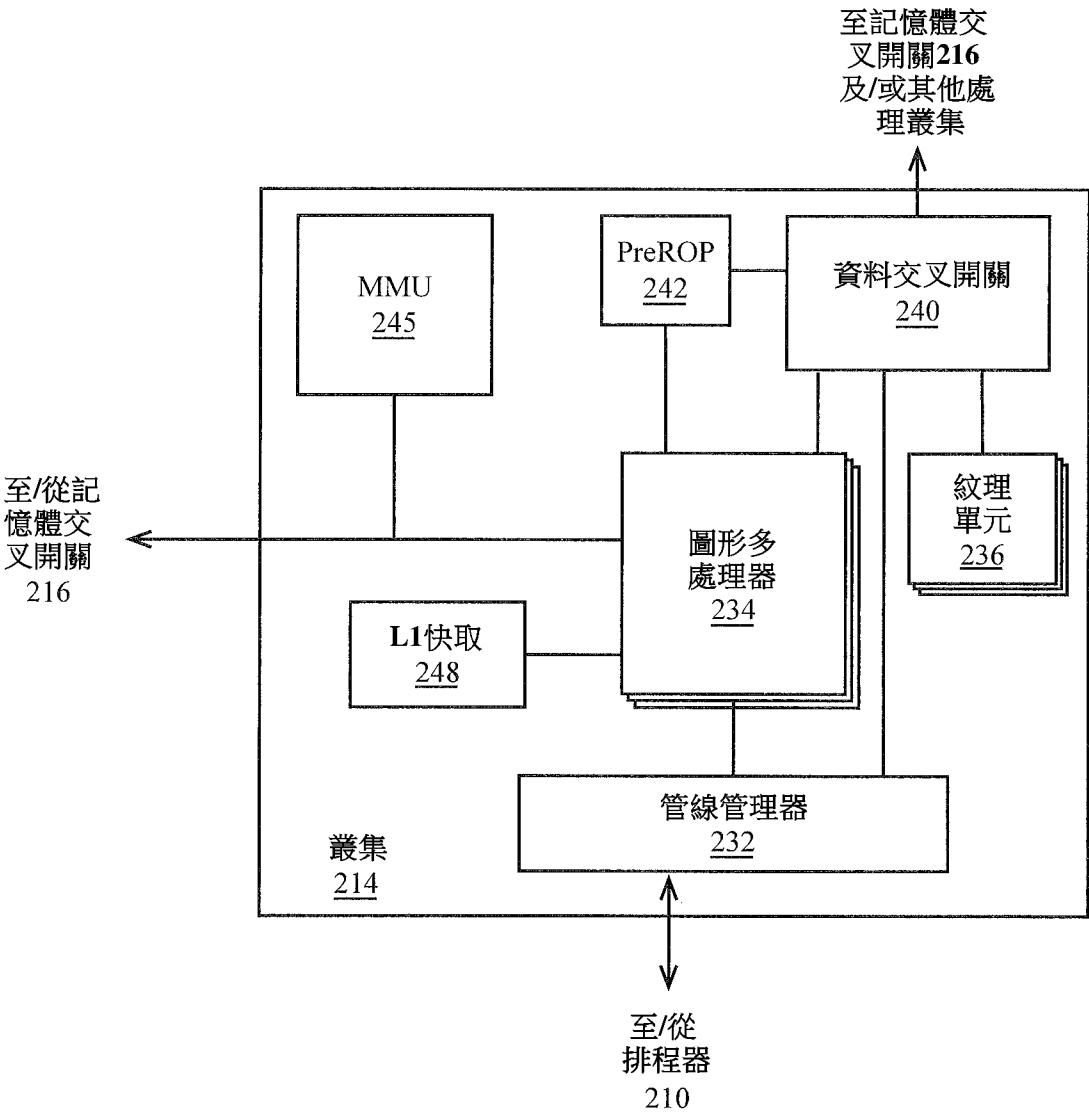
【圖 1】



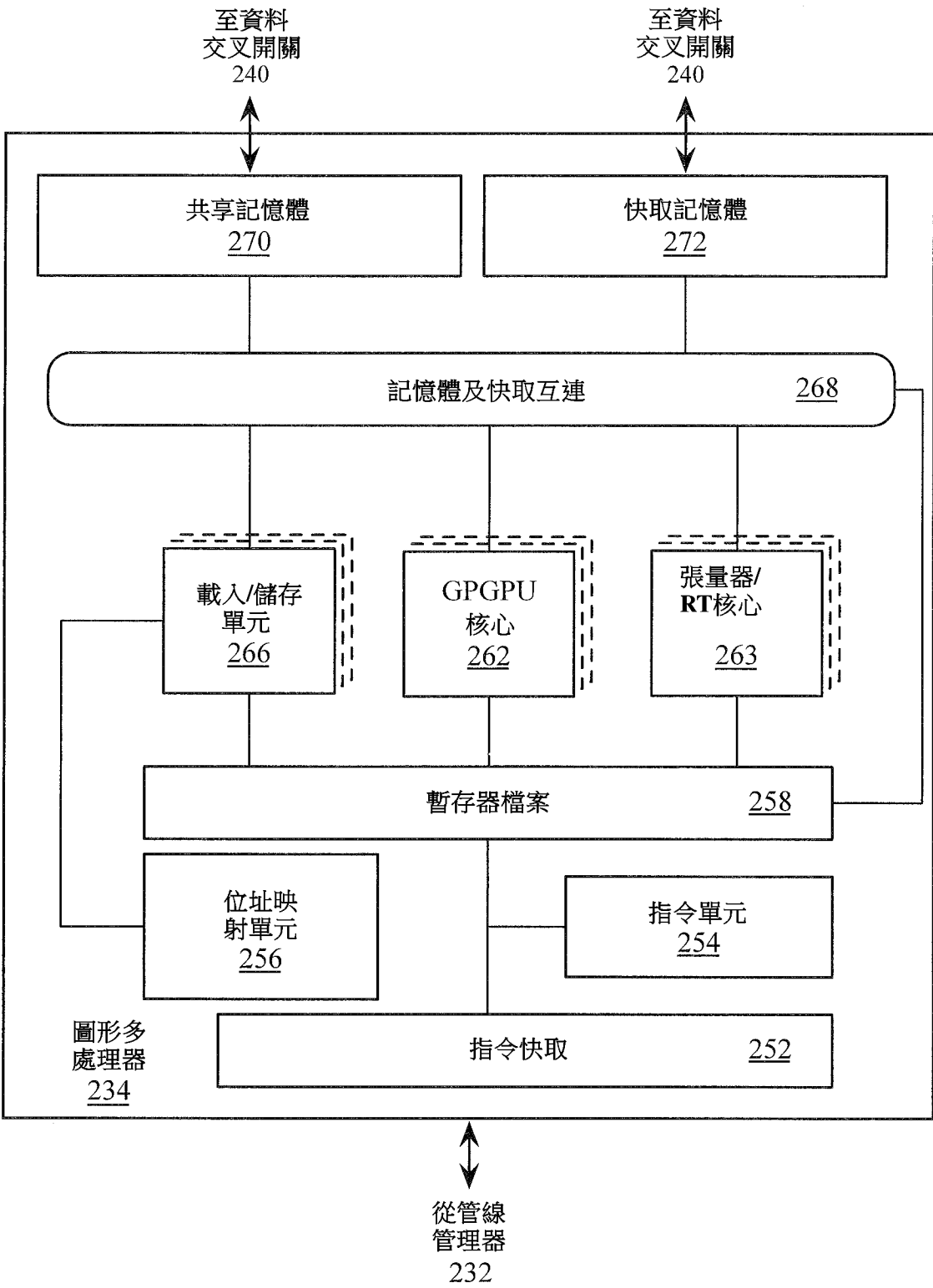
【圖 2A】



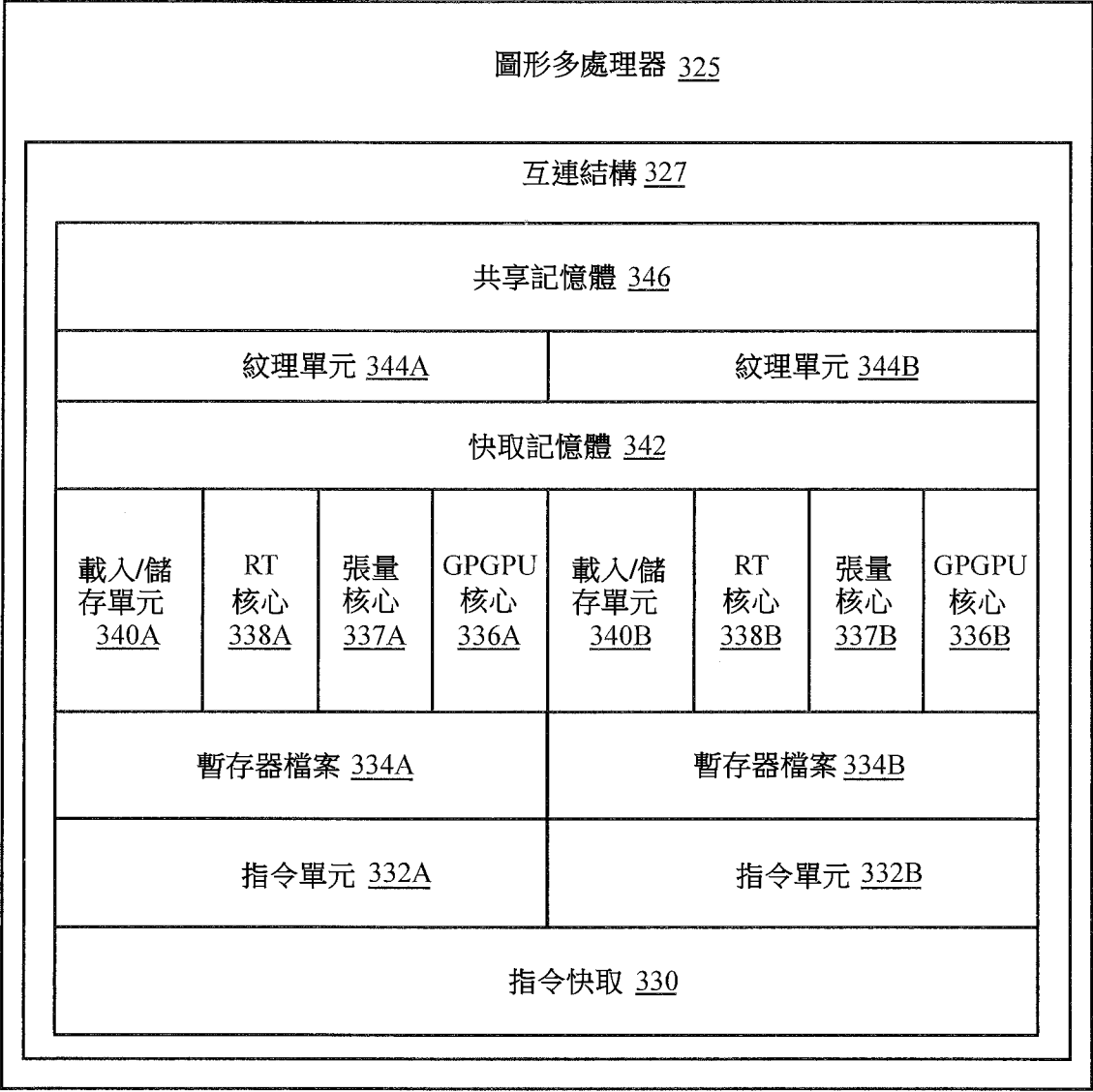
【圖 2B】



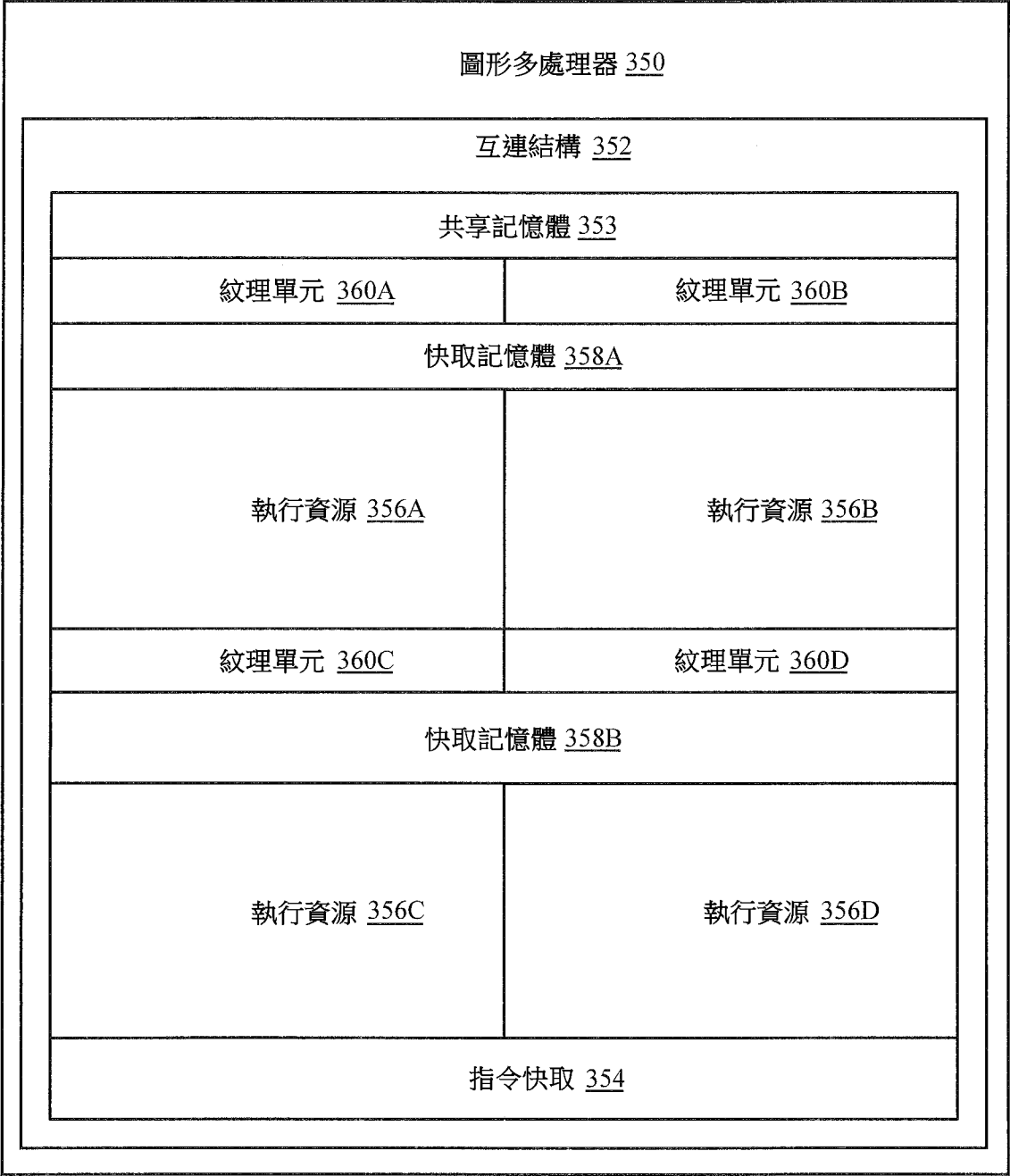
【圖 2C】



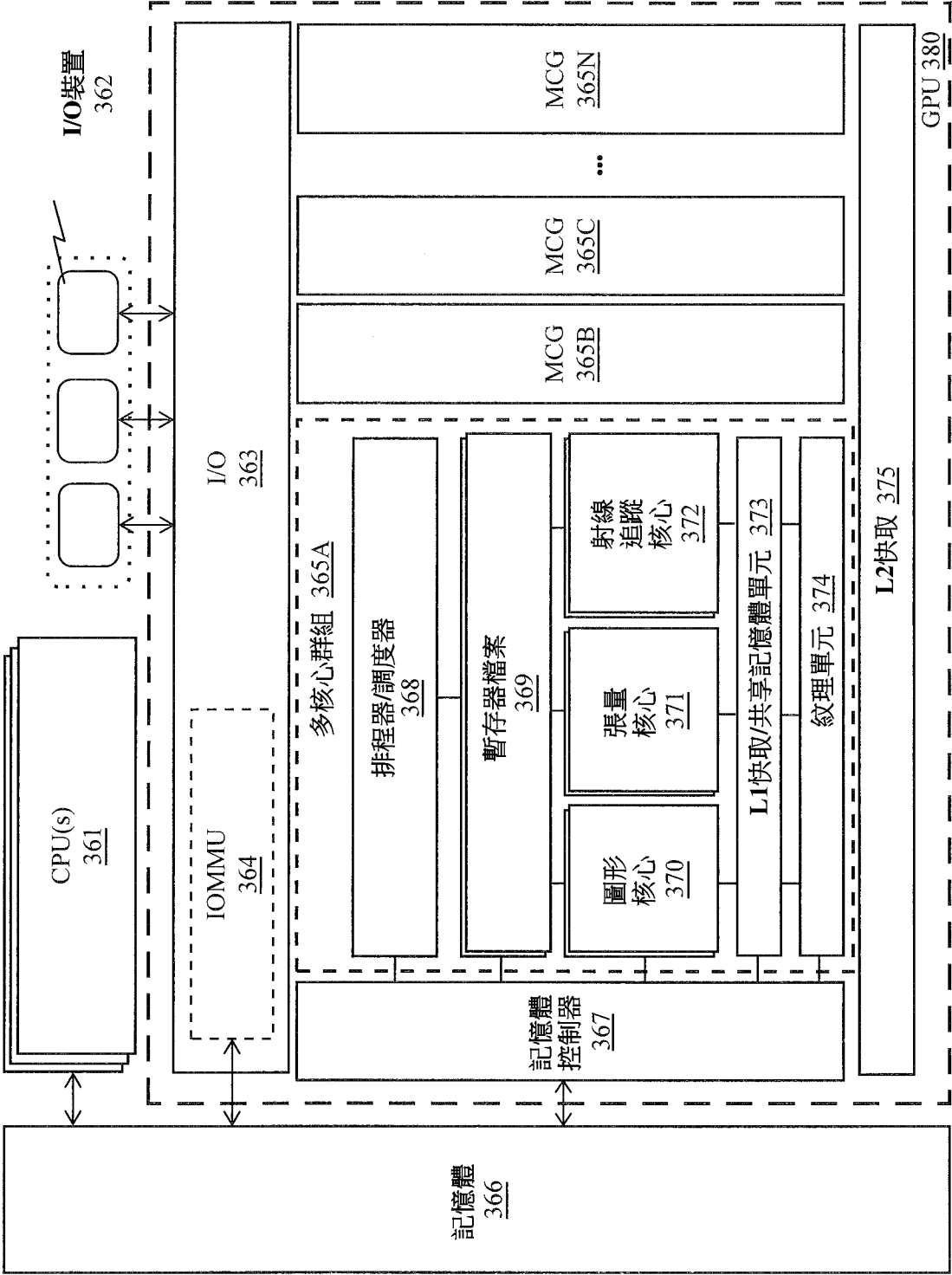
【圖 2D】



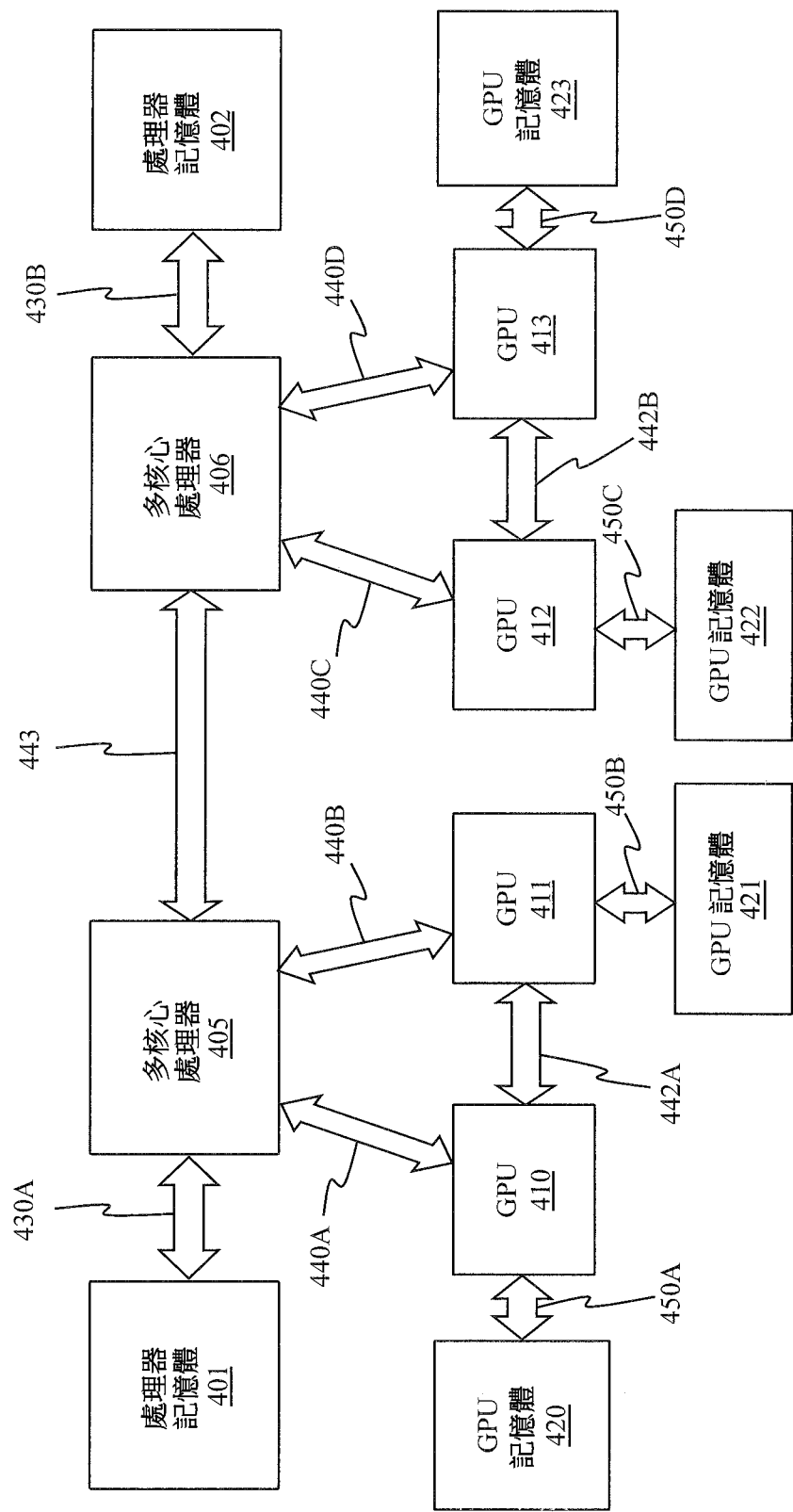
【圖 3A】



【圖 3B】

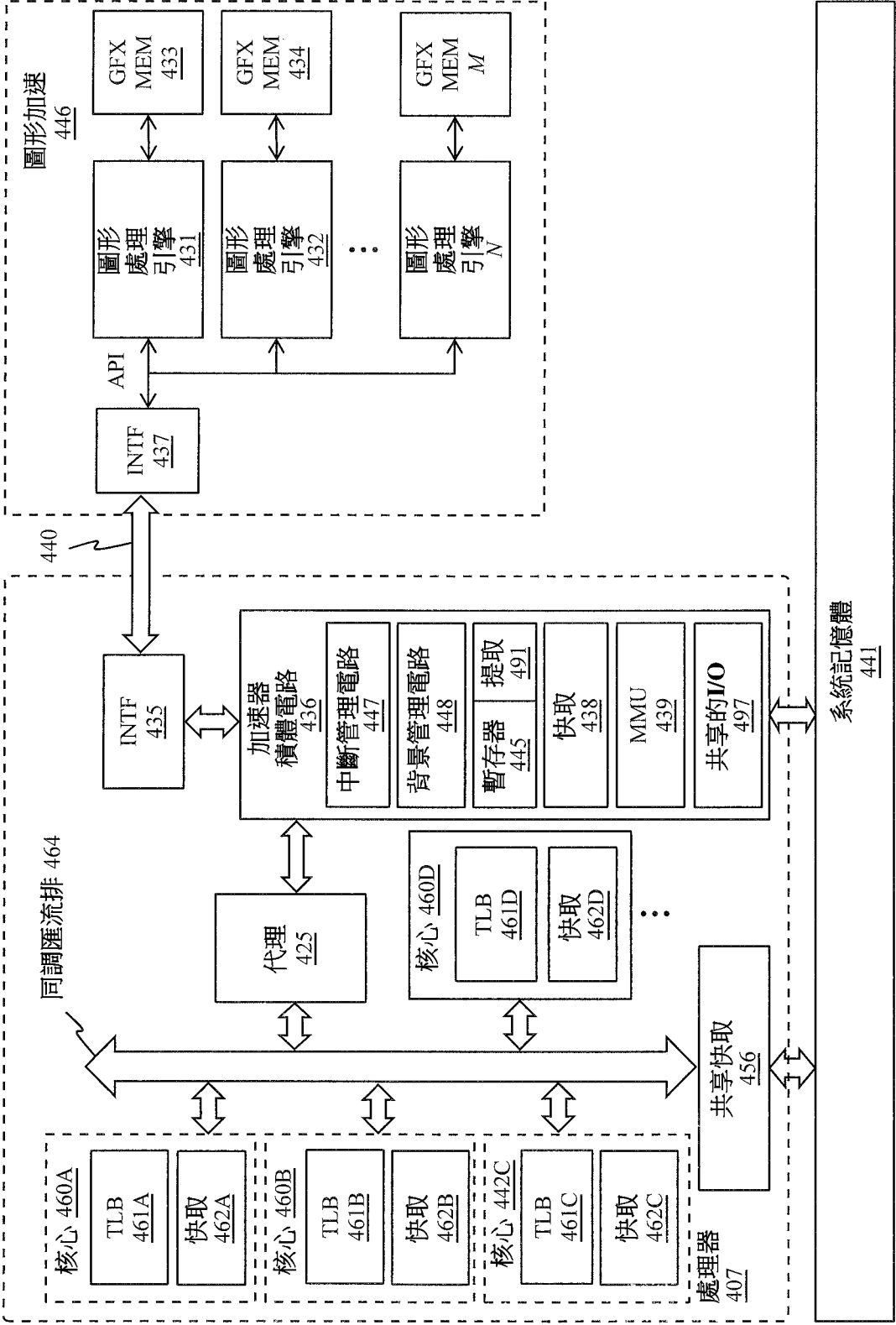


【圖 3C】

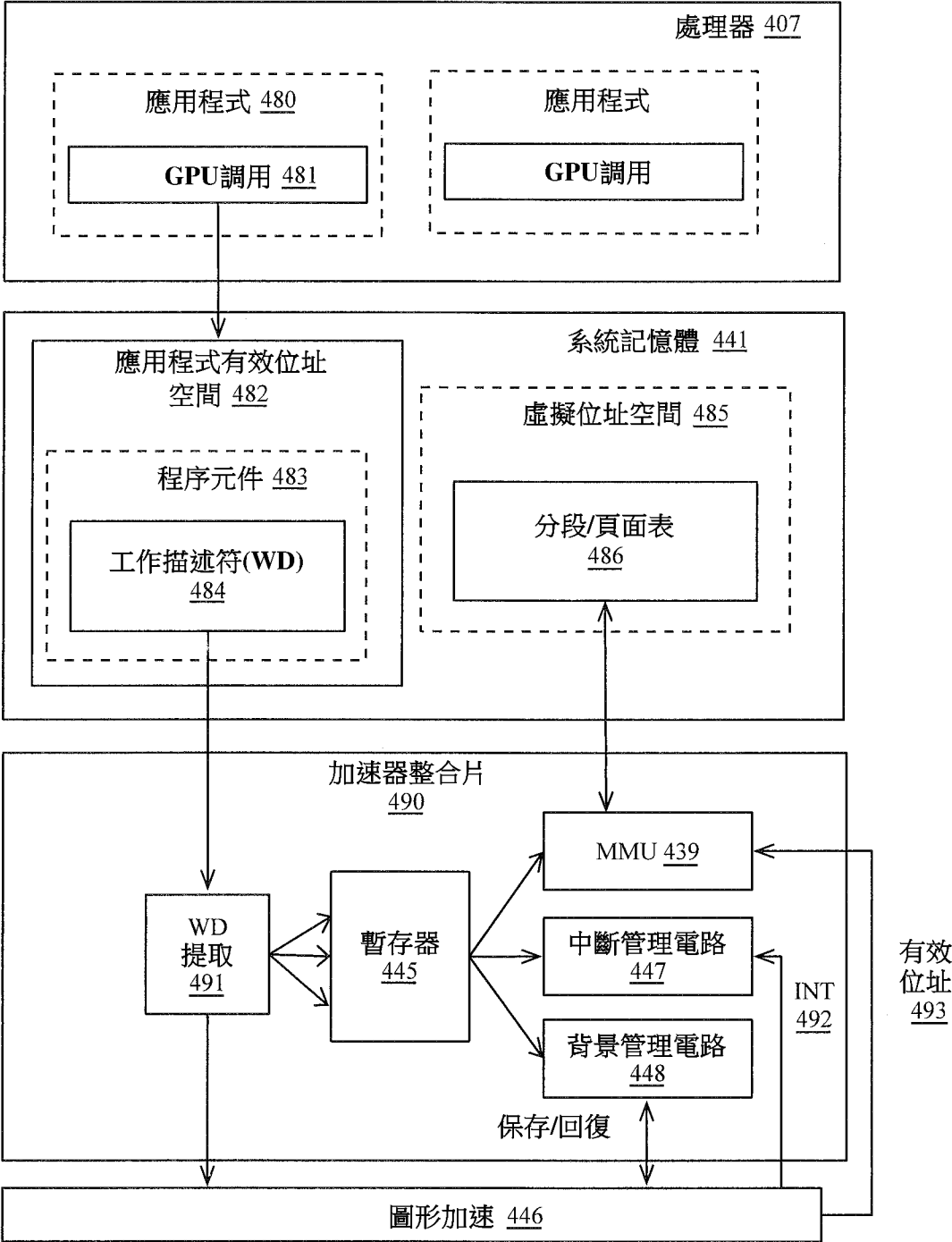


【圖 4A】

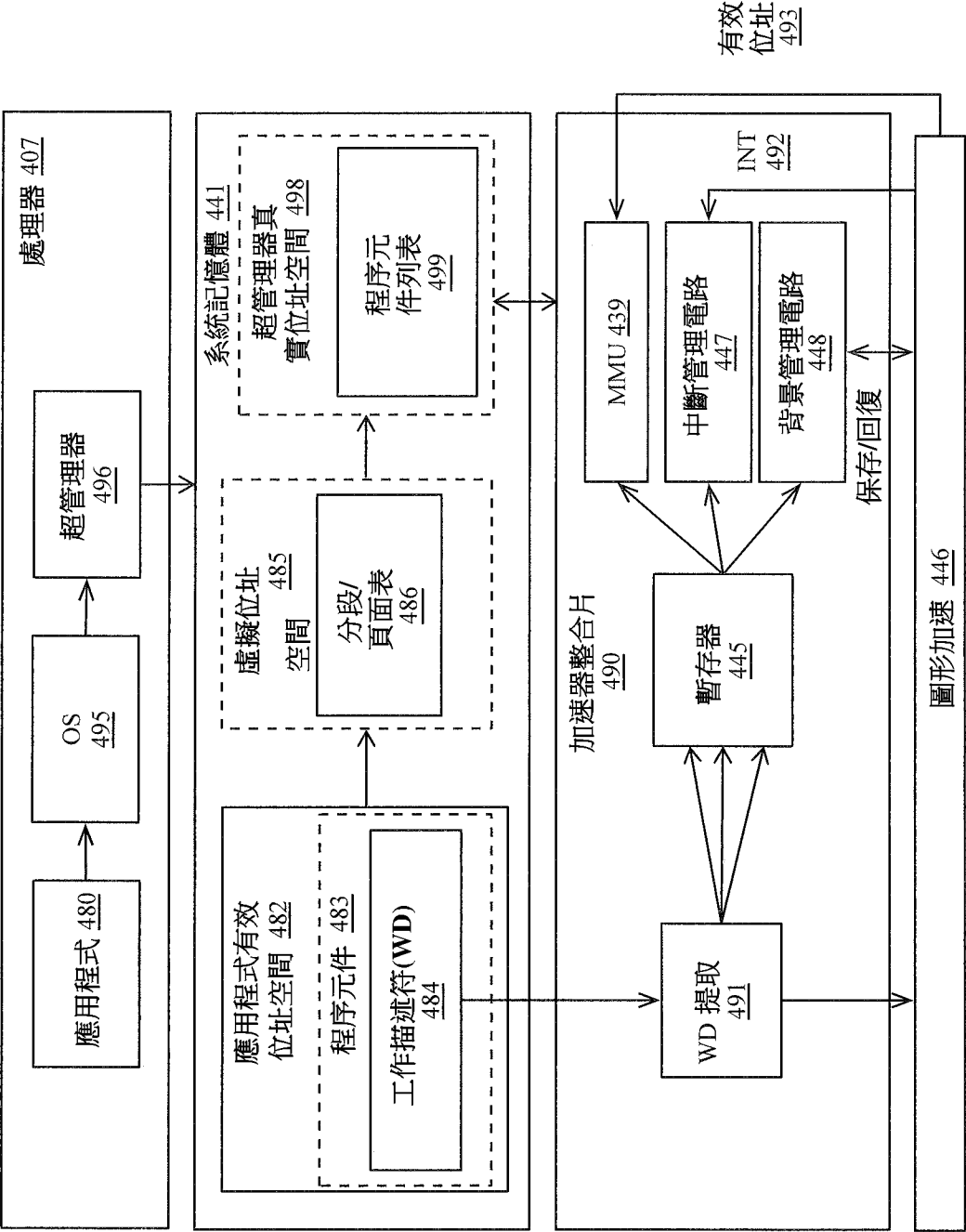




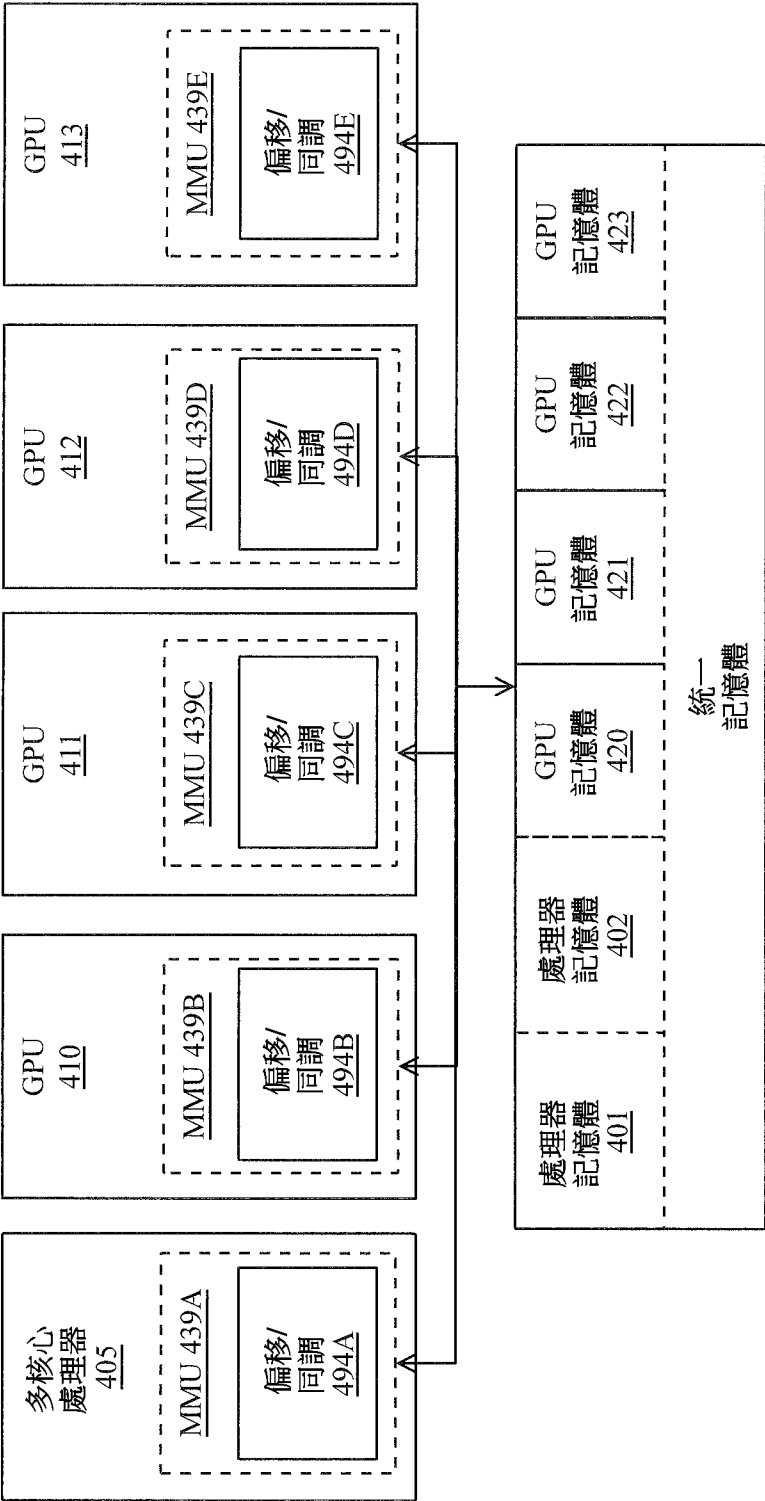
【圖 4C】



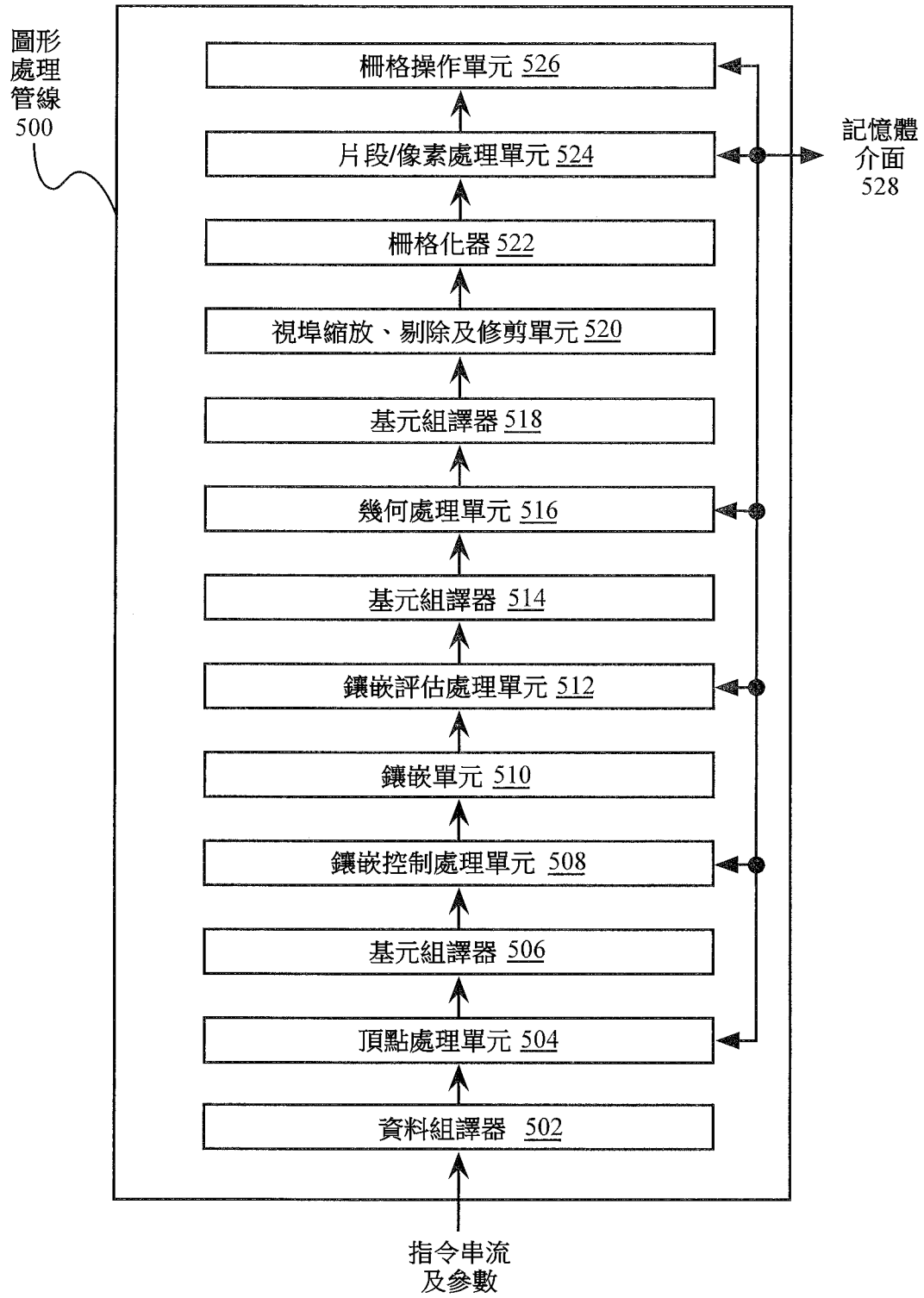
【圖 4D】



【圖 4E】

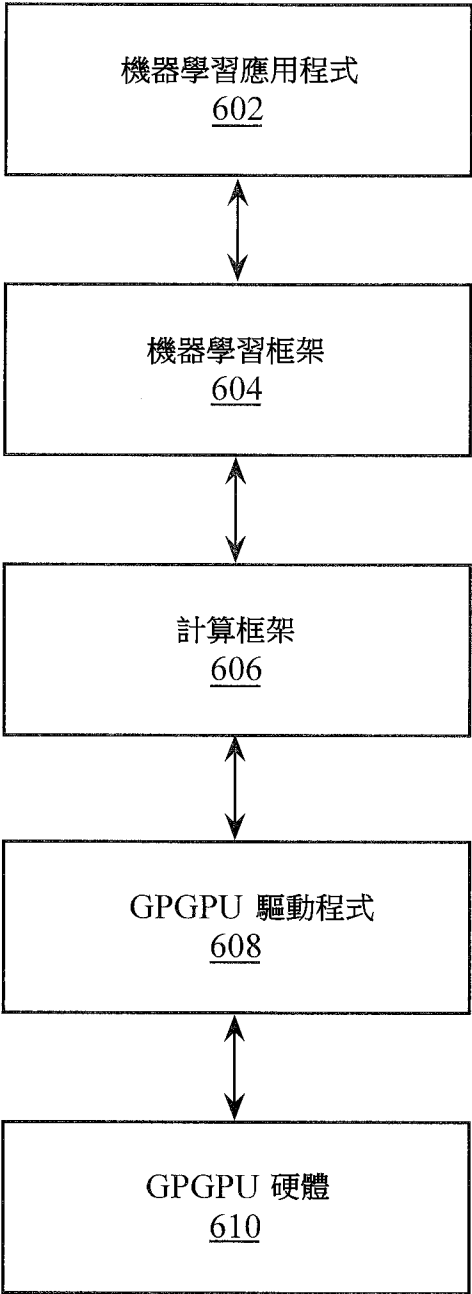


【圖 4F】

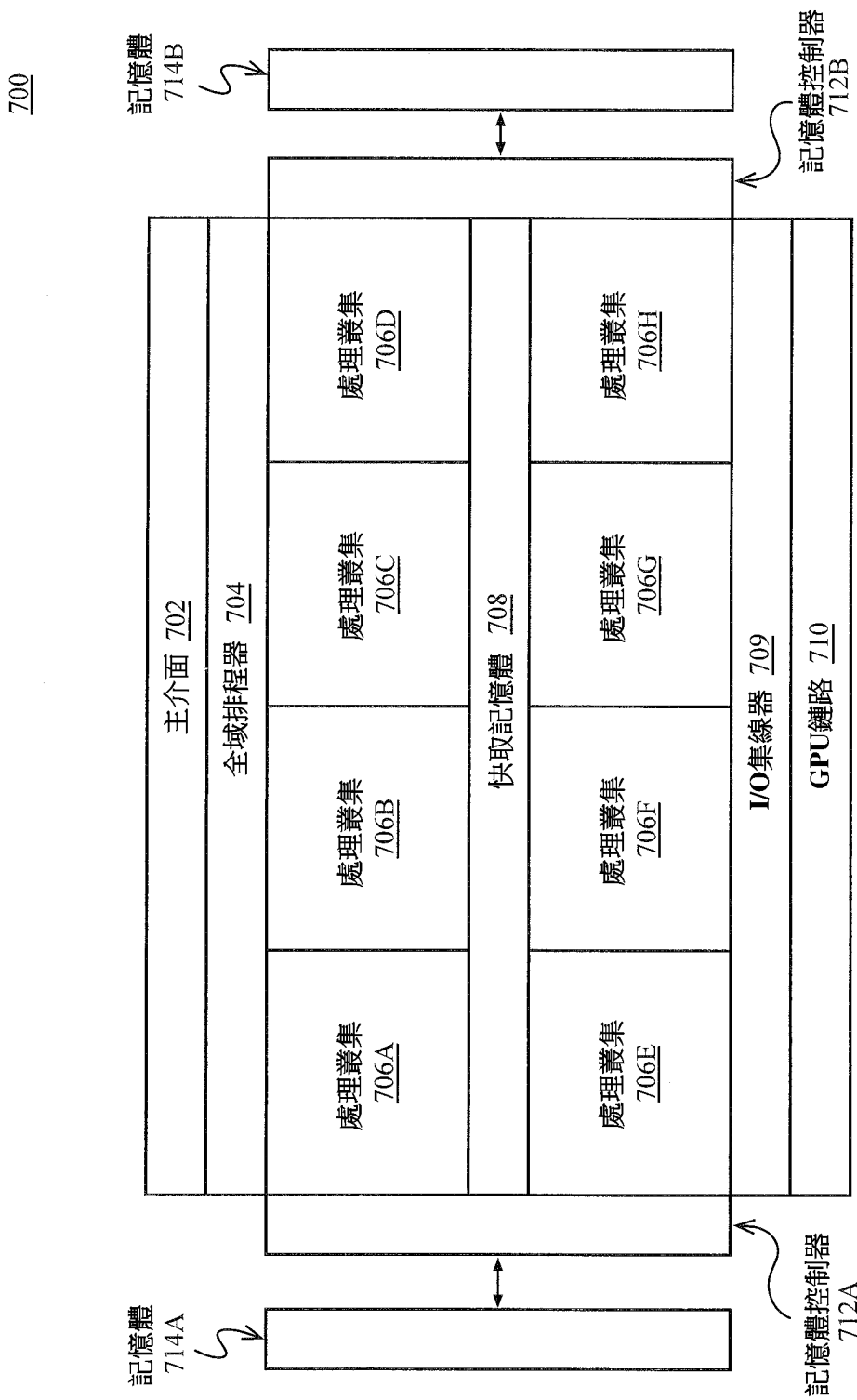


【圖 5】

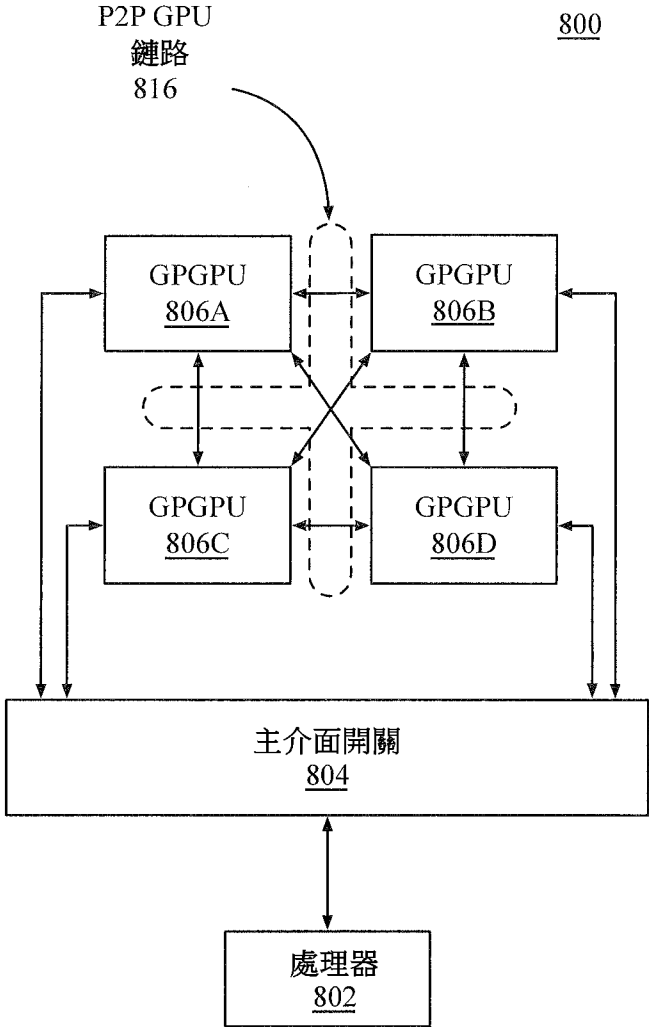
600



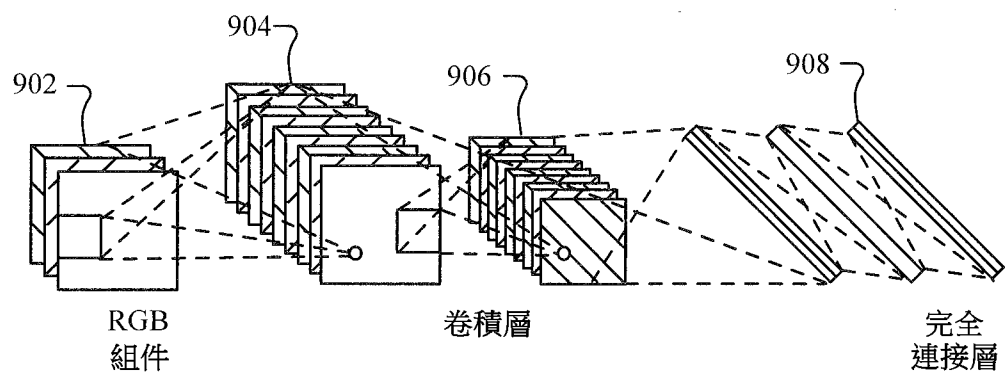
【圖 6】



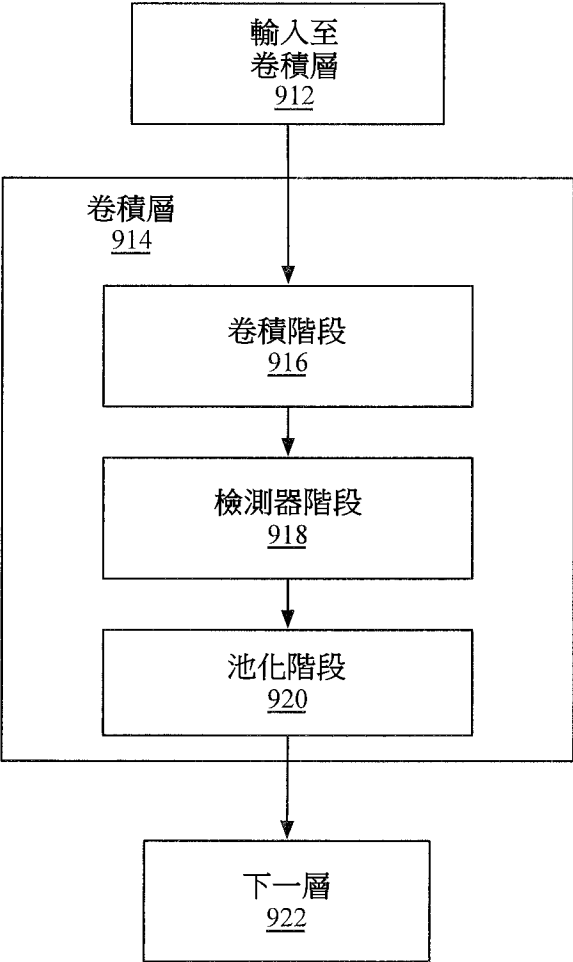
【圖 7】



【圖 8】

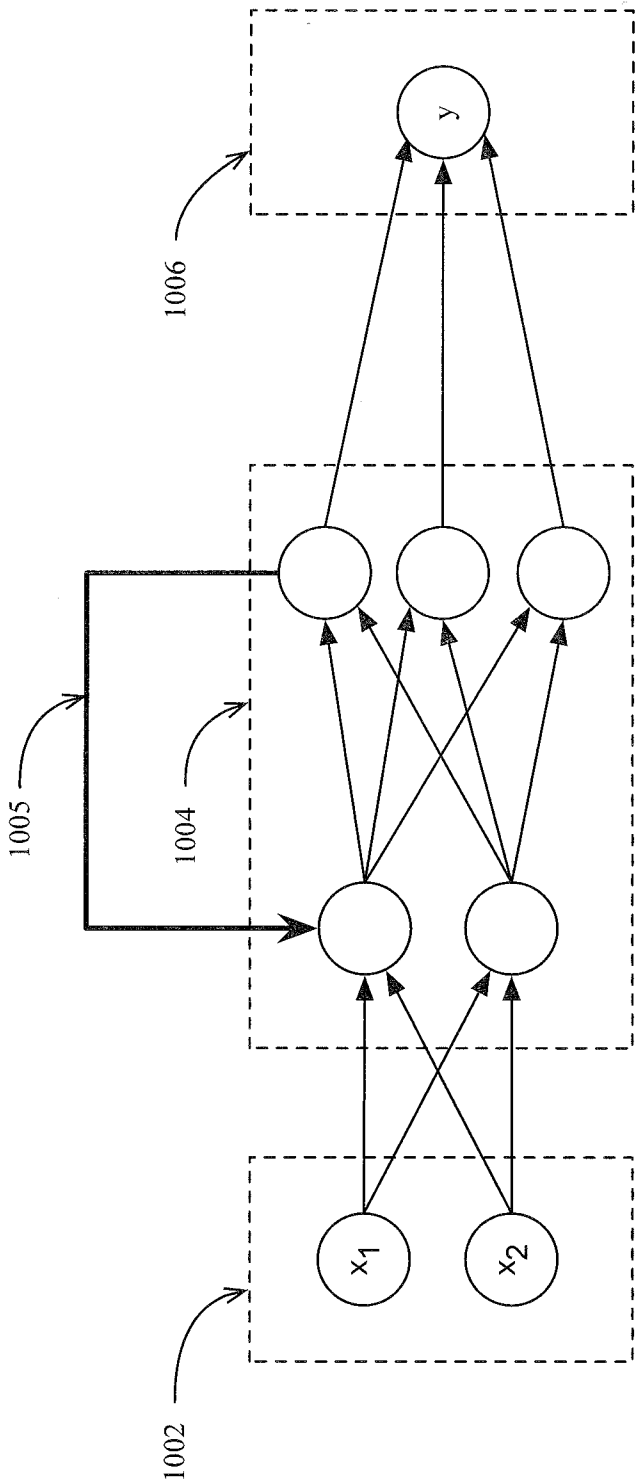


【圖 9A】

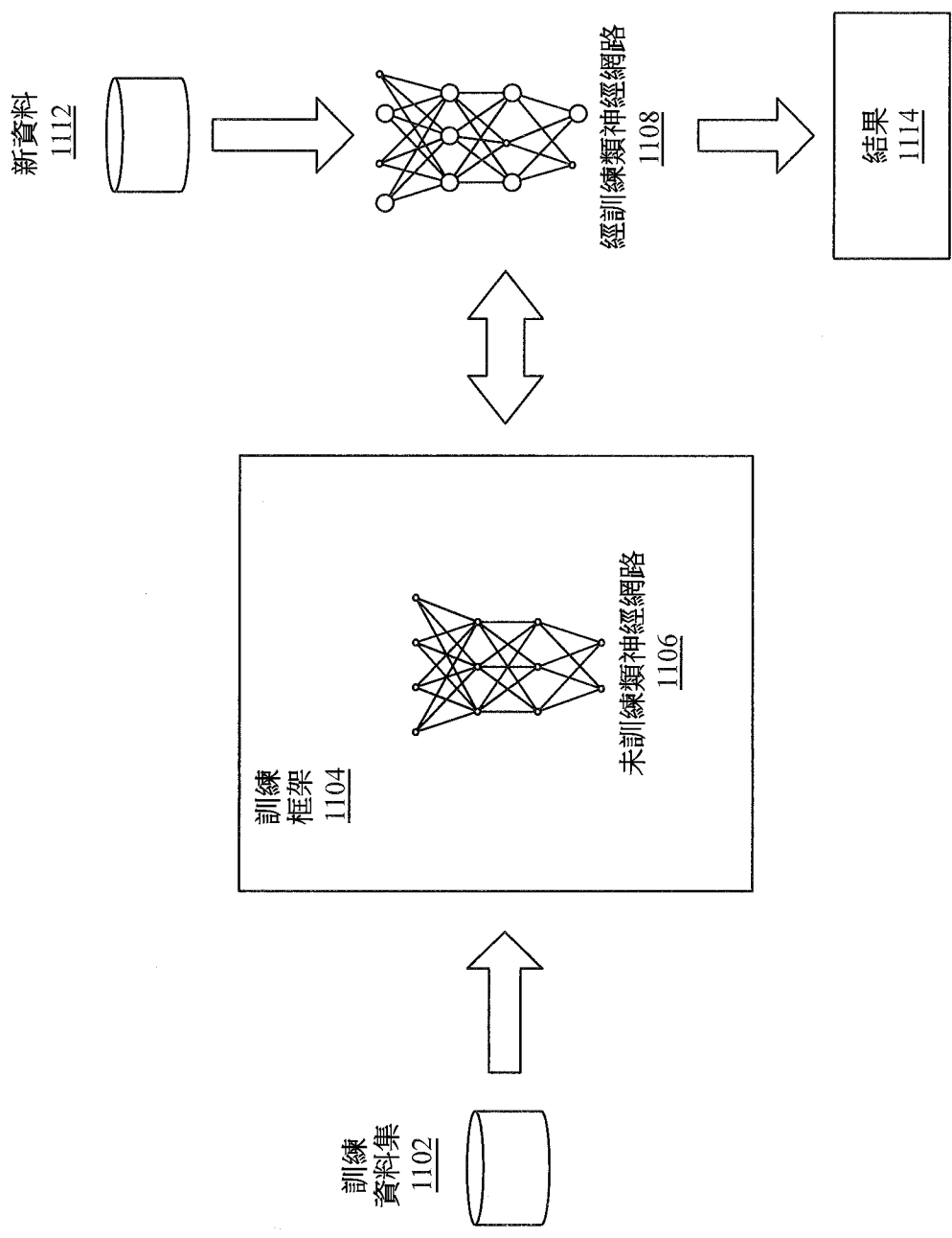


【圖 9B】

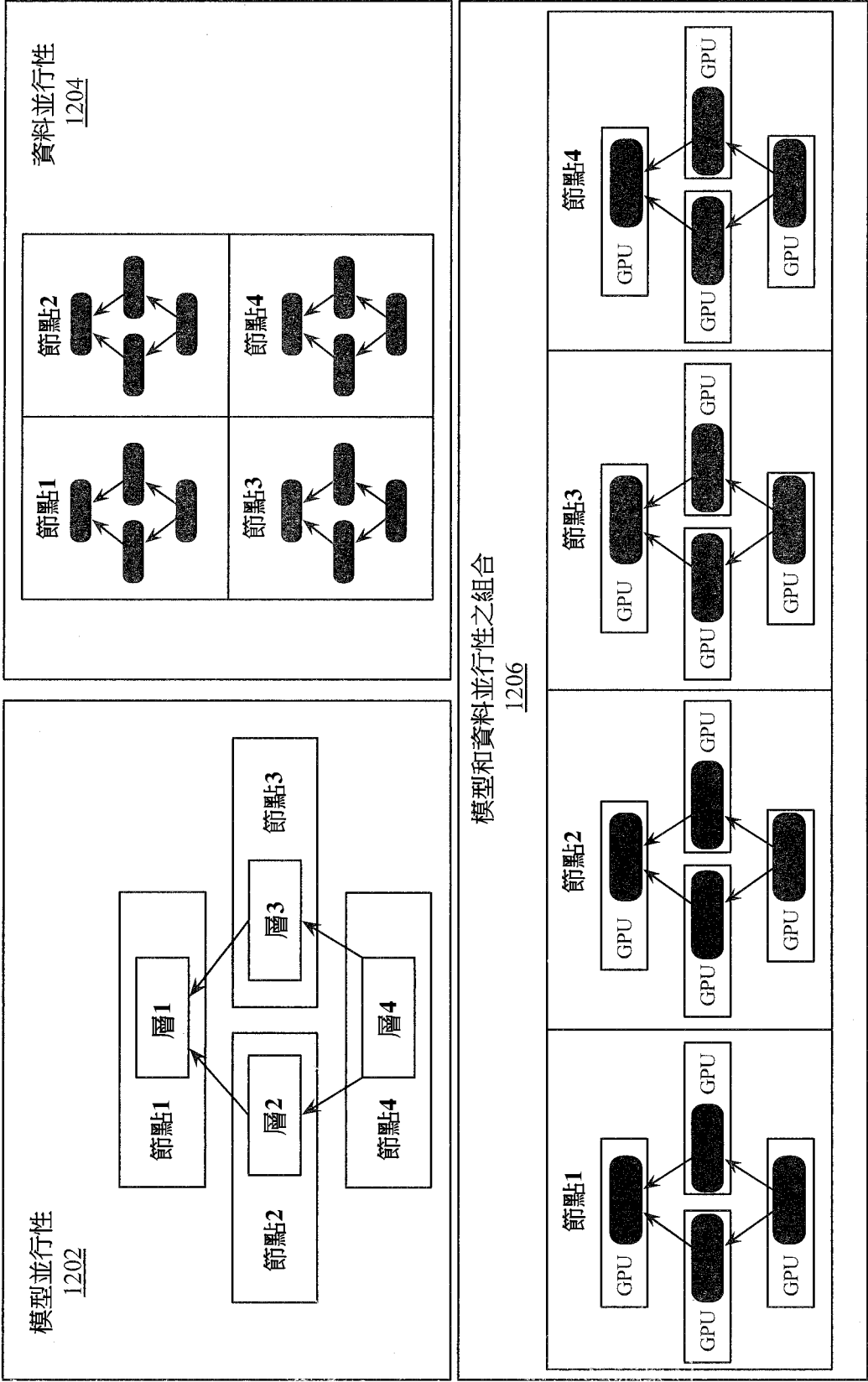
1000



【圖 10】

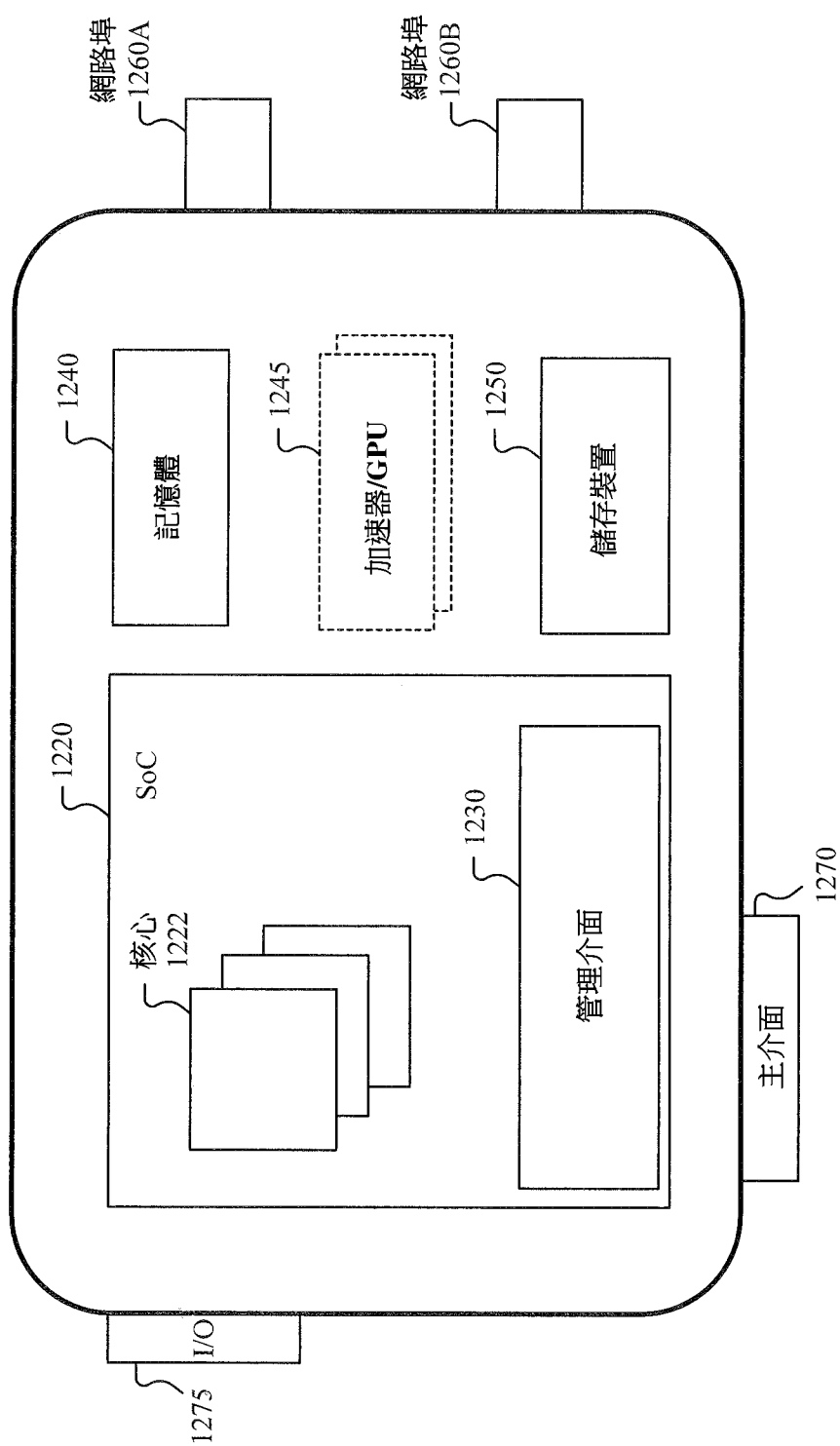


【圖 11】



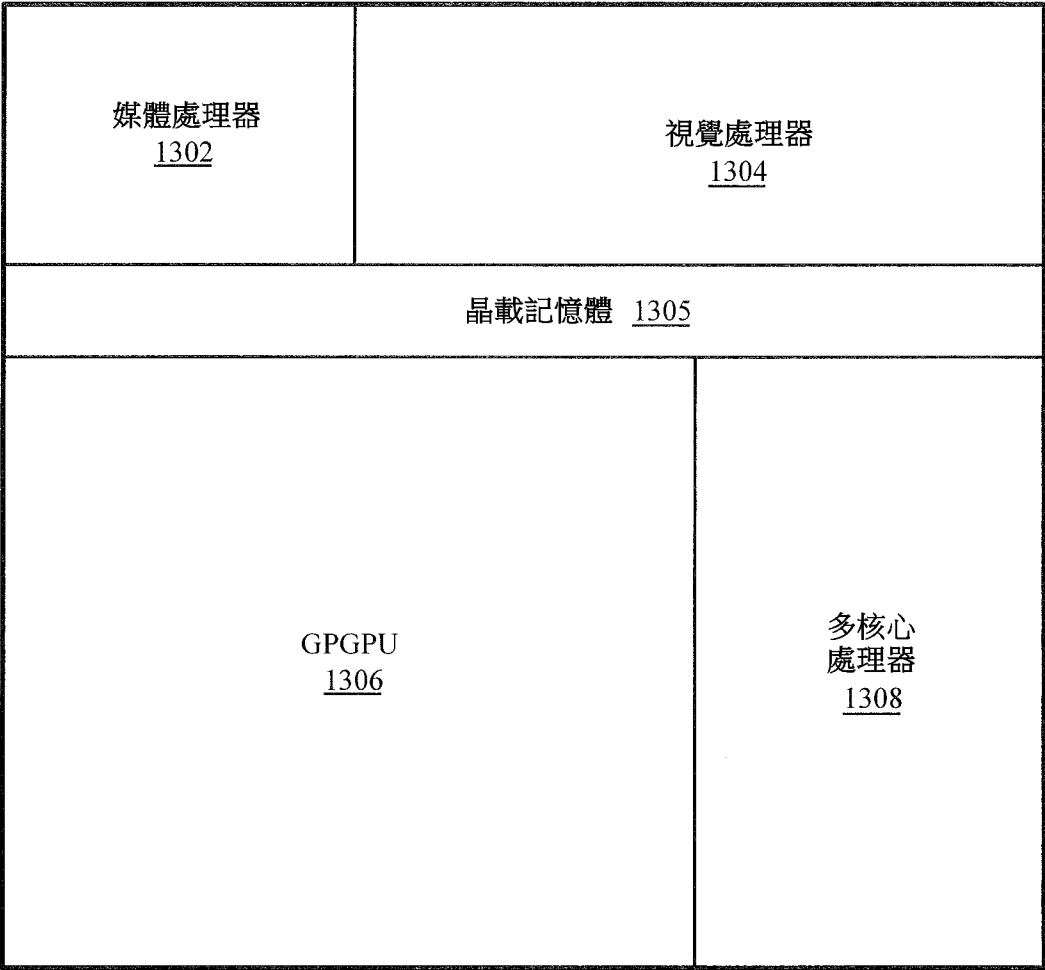
【圖 12A】

1210

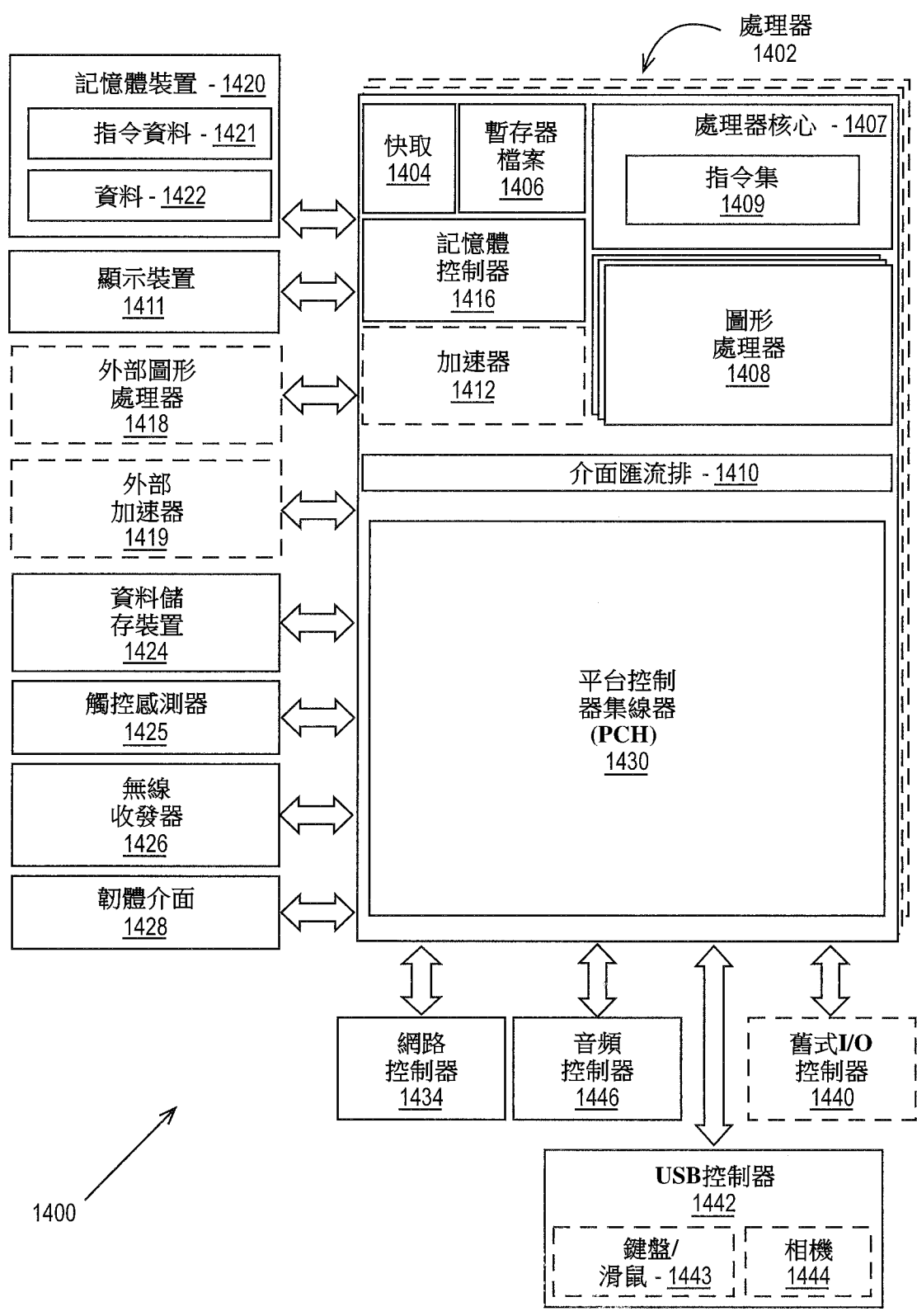


【圖 12B】

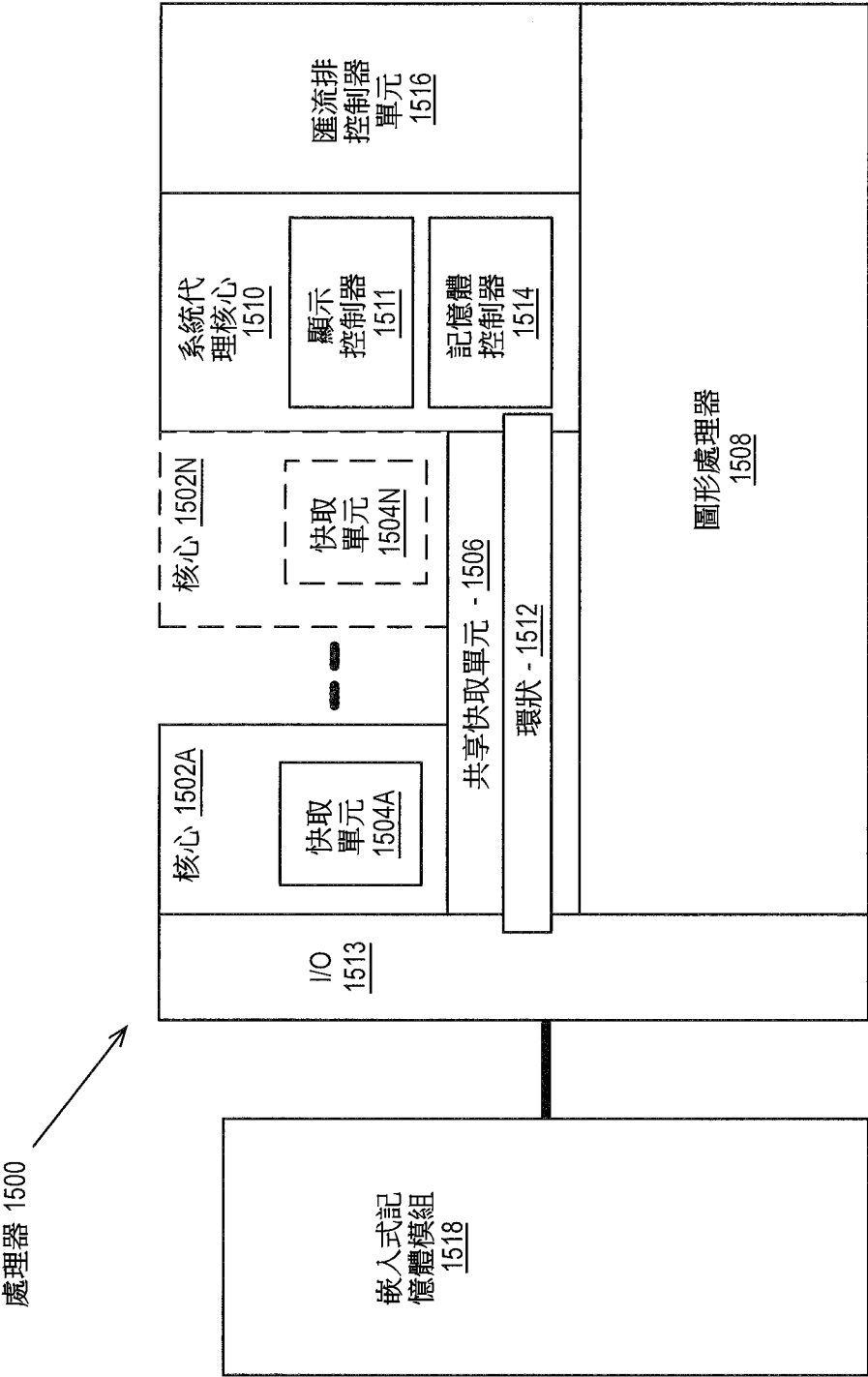
1300



【圖 13】

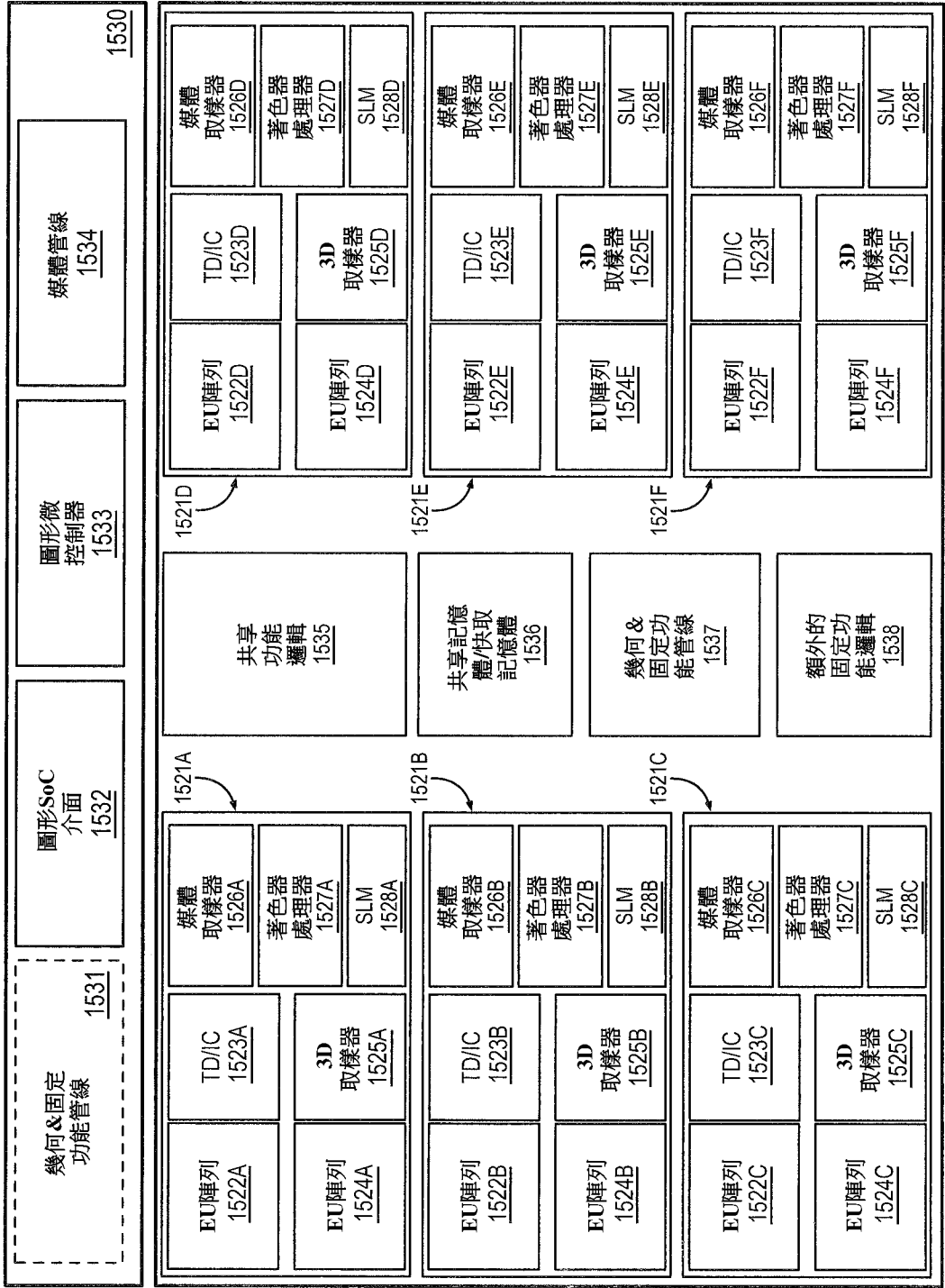


【圖 14】

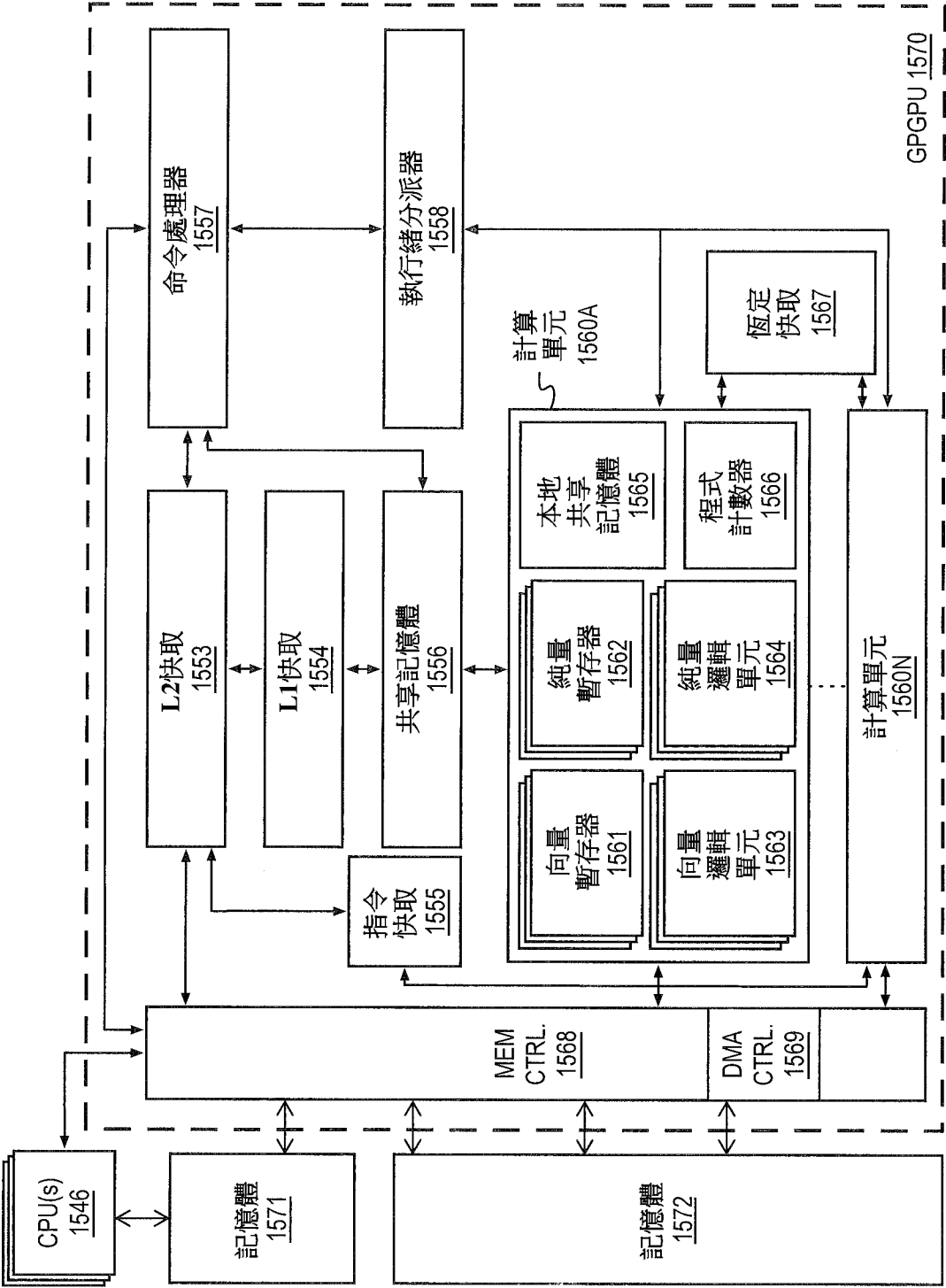


【圖 15A】

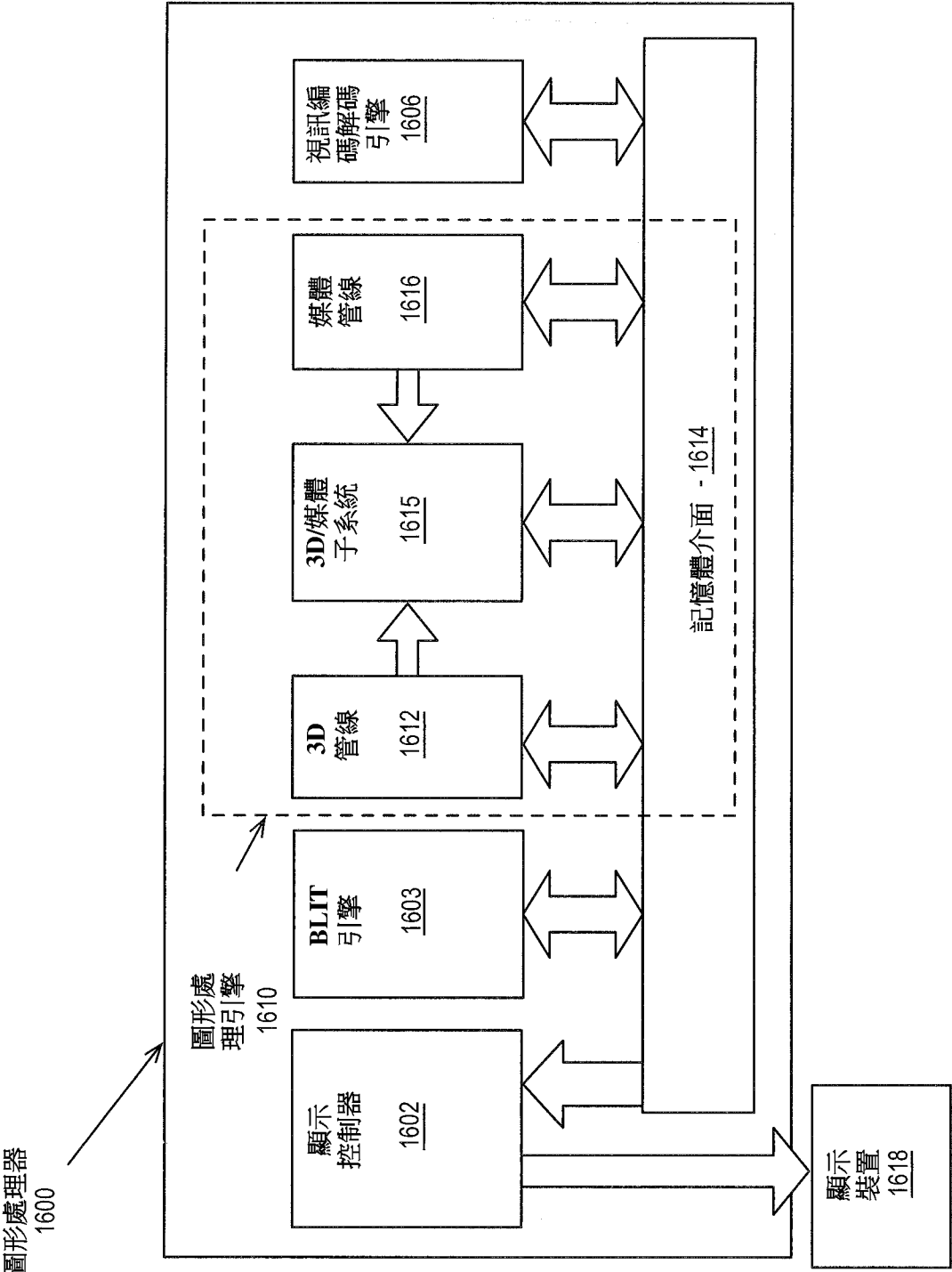
1519



【圖 15B】

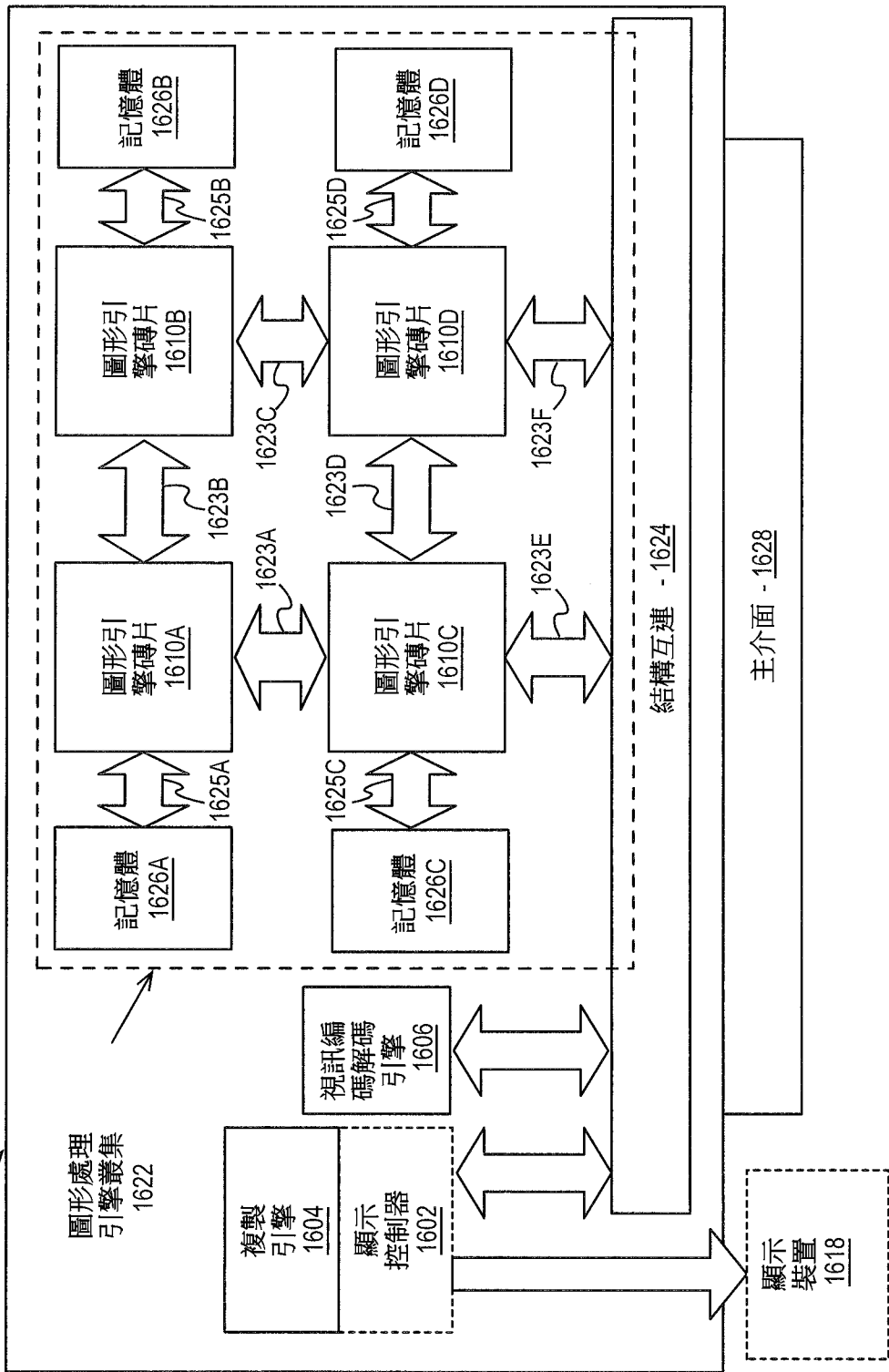


【圖 15C】

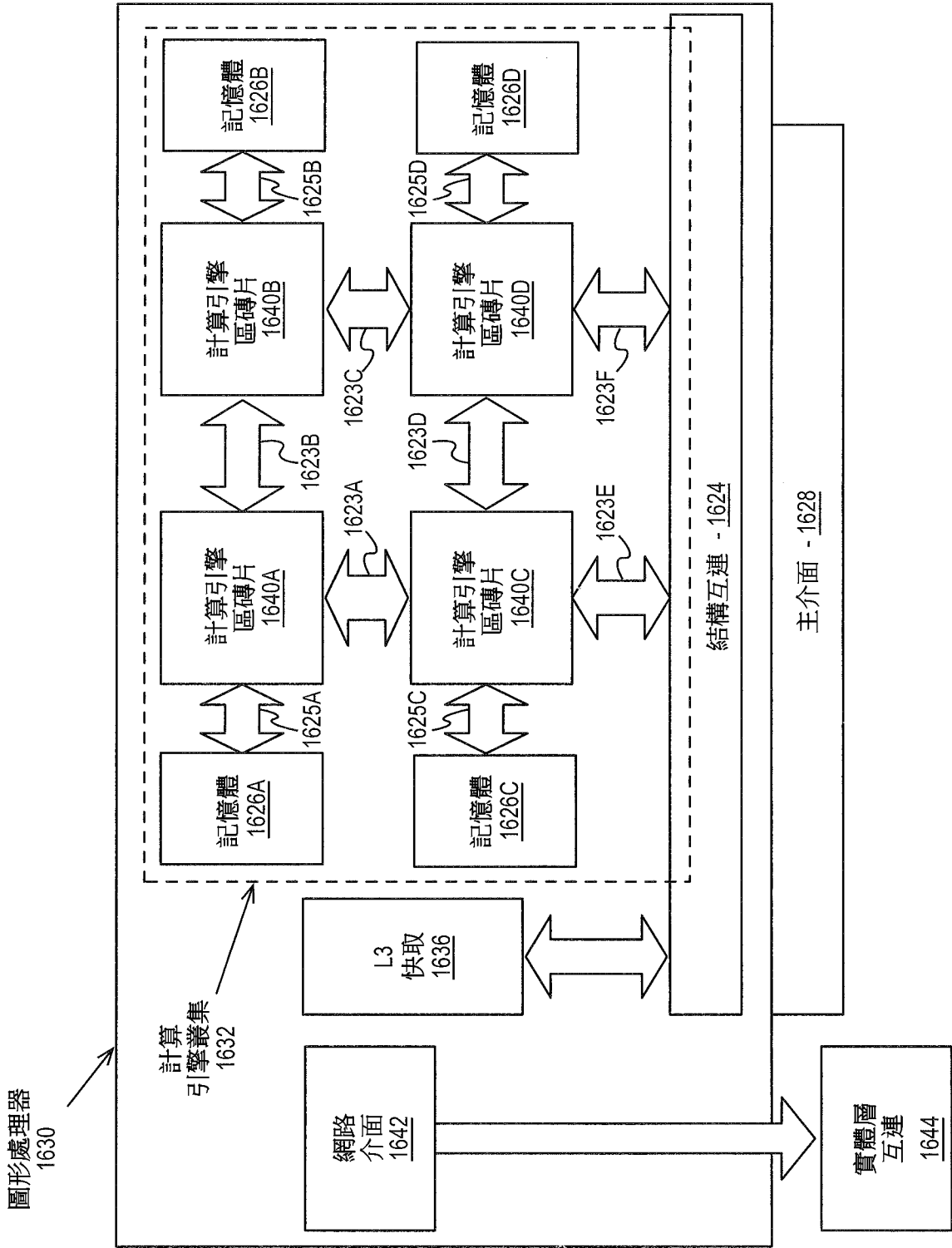


【圖 16A】

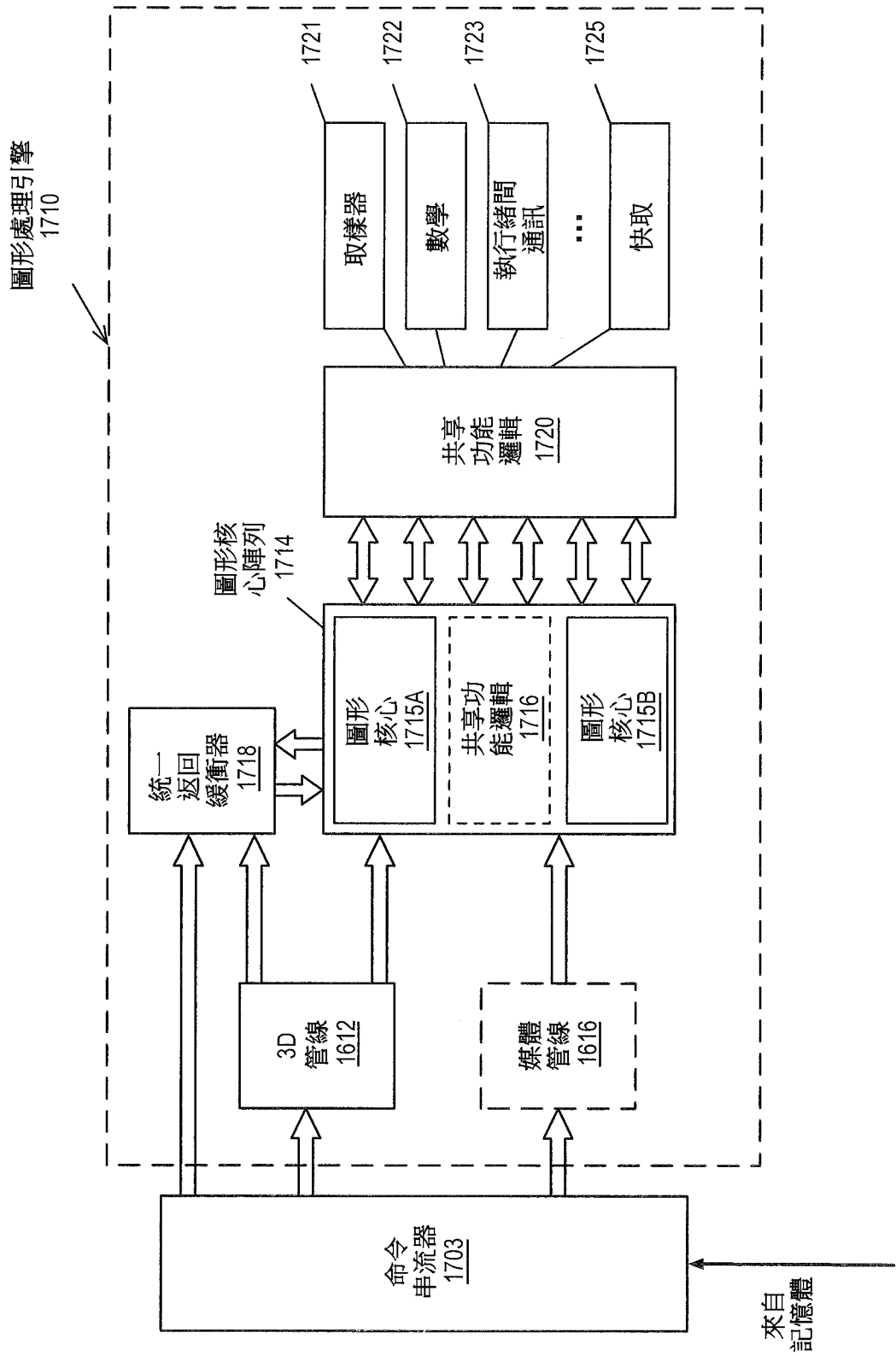
圖形處理器
1620



【圖 16B】

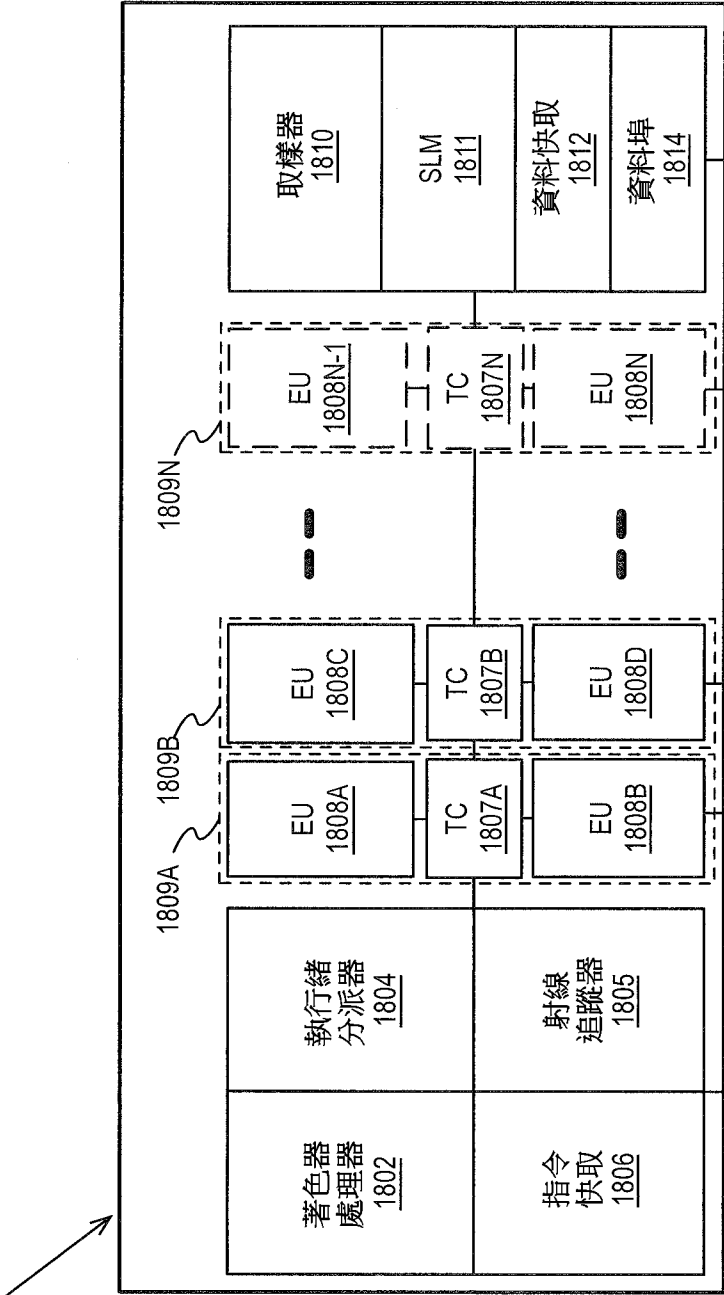


【圖 16C】

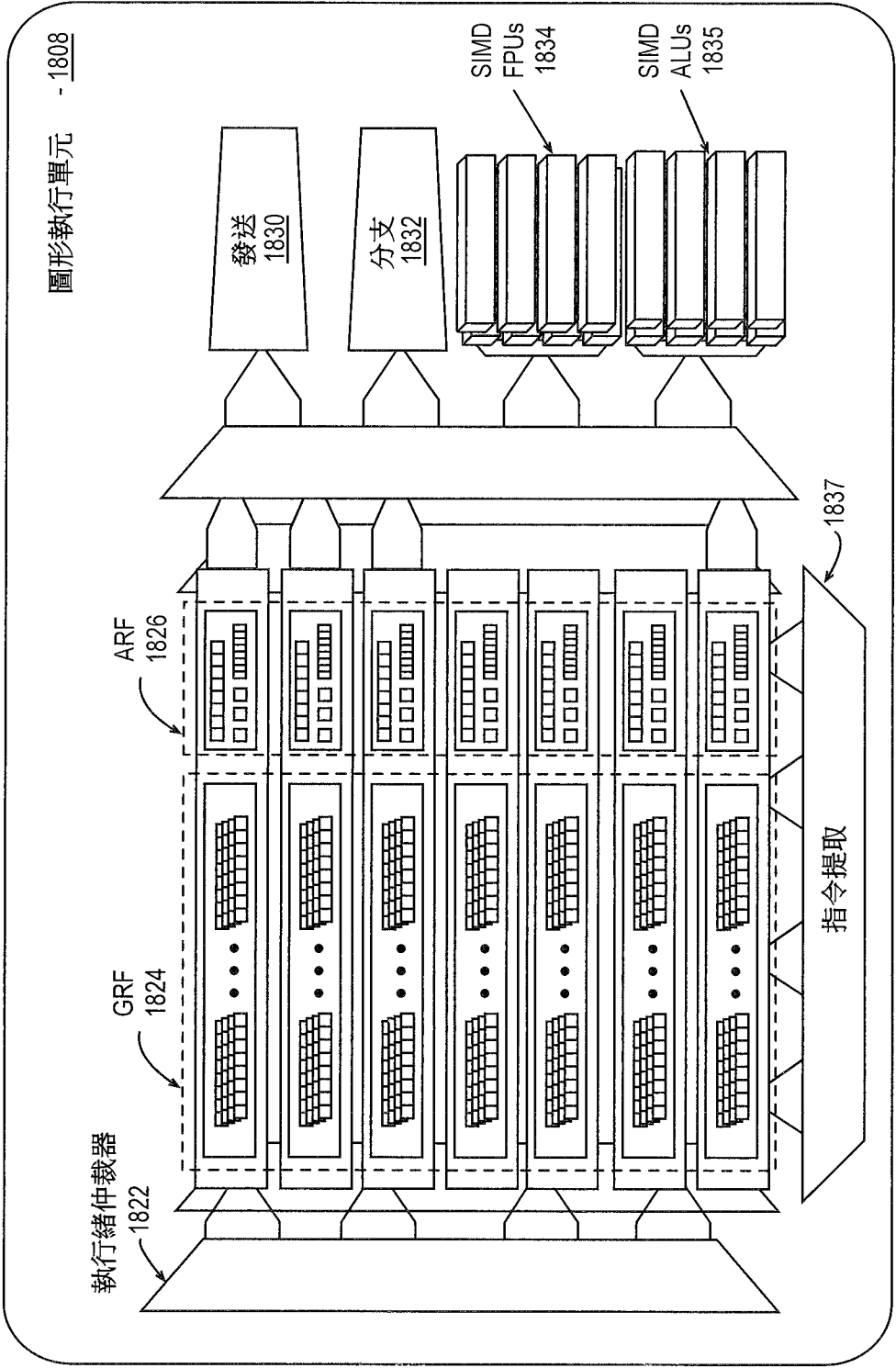


【圖 17】

執行邏輯
1800

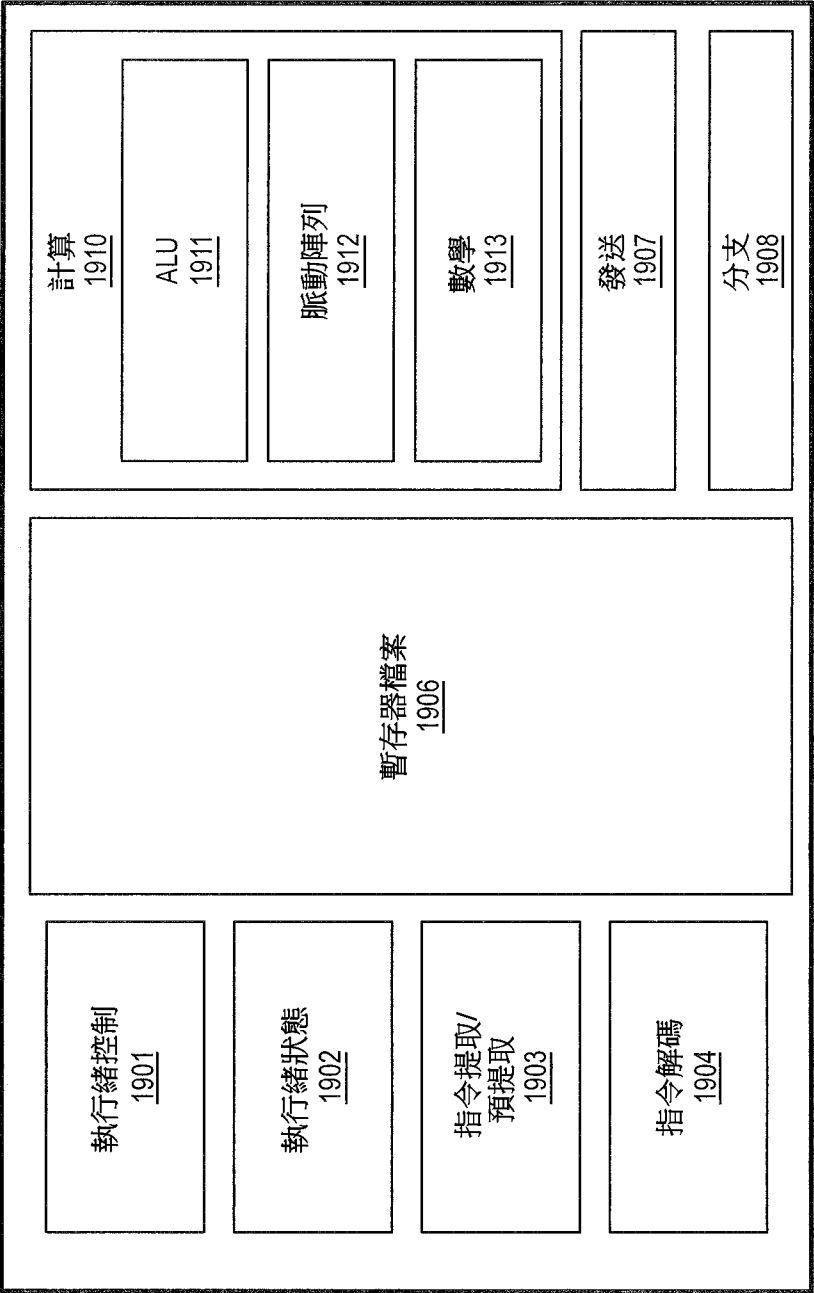


【圖 18A】



【圖 18B】

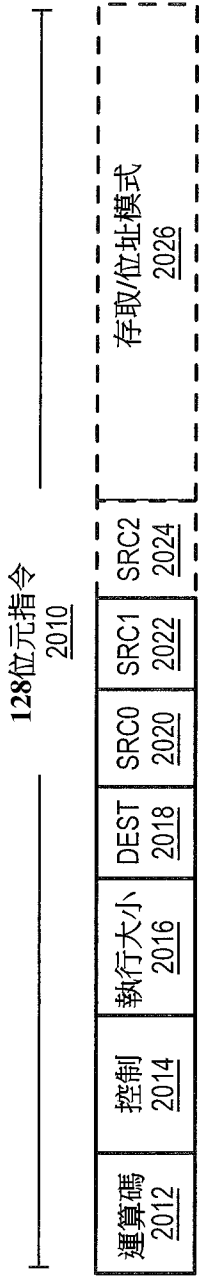
執行單元
1900



【圖 19】

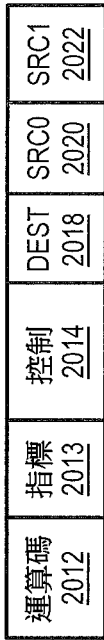
圖形處理器指令格式

2000



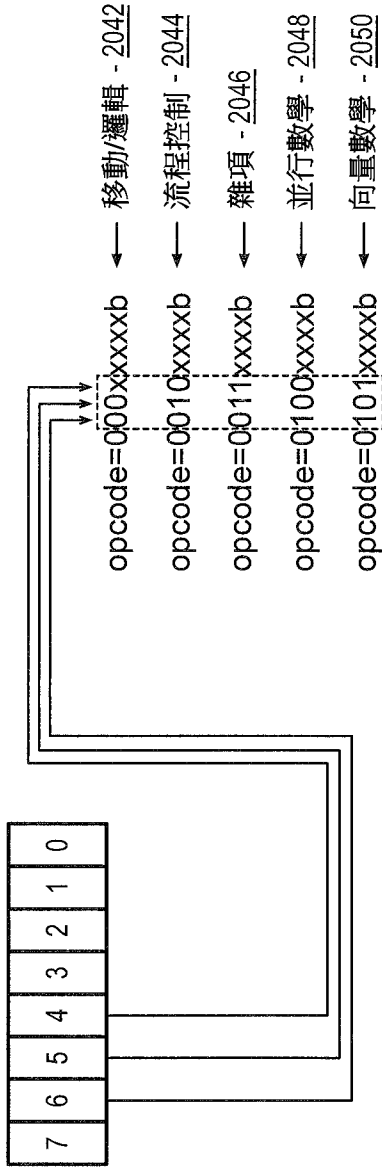
64位元緊湊指令

2030

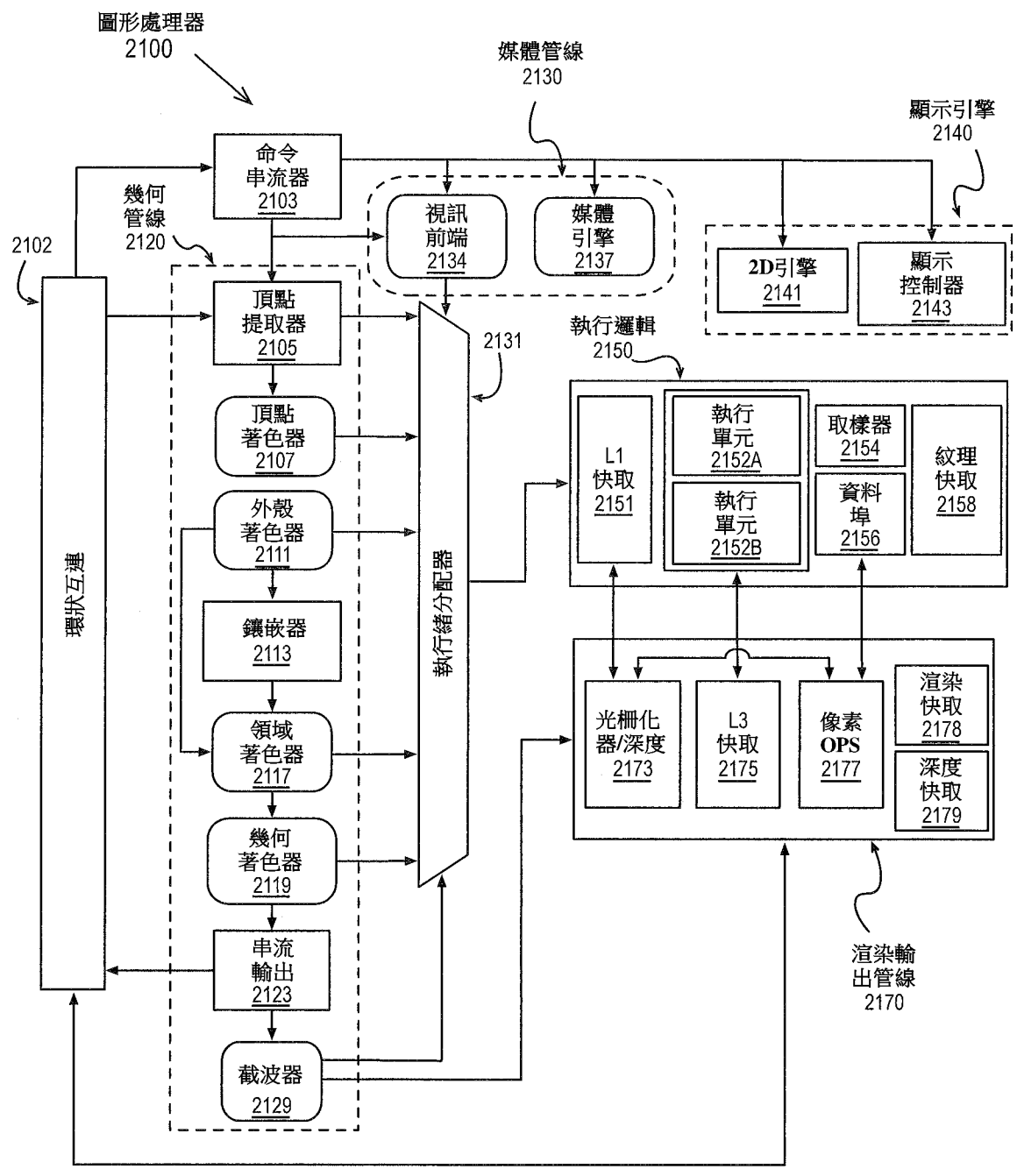


運算碼解碼

2040



【圖 20】



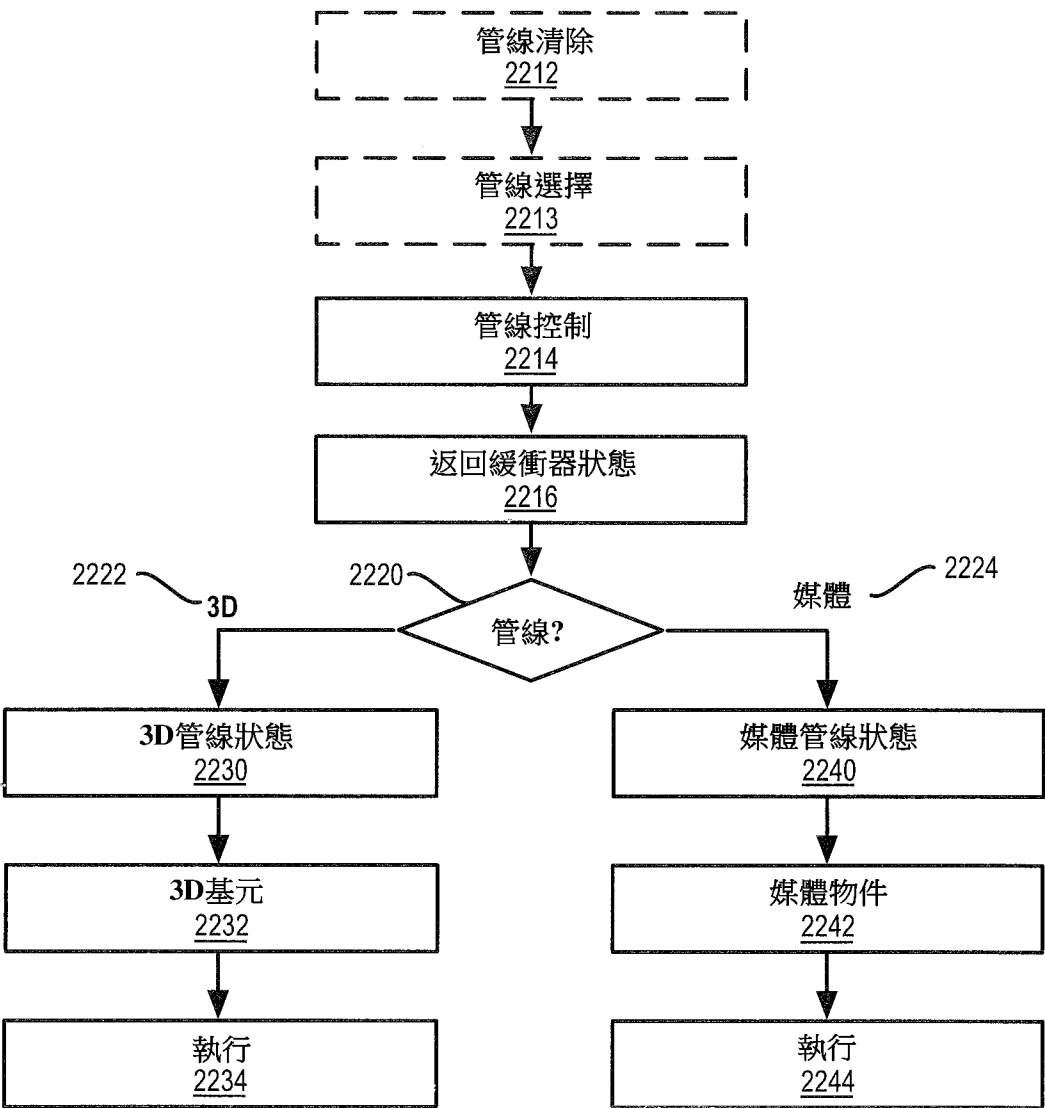
【圖 21】

圖形處理器命令格式
2200



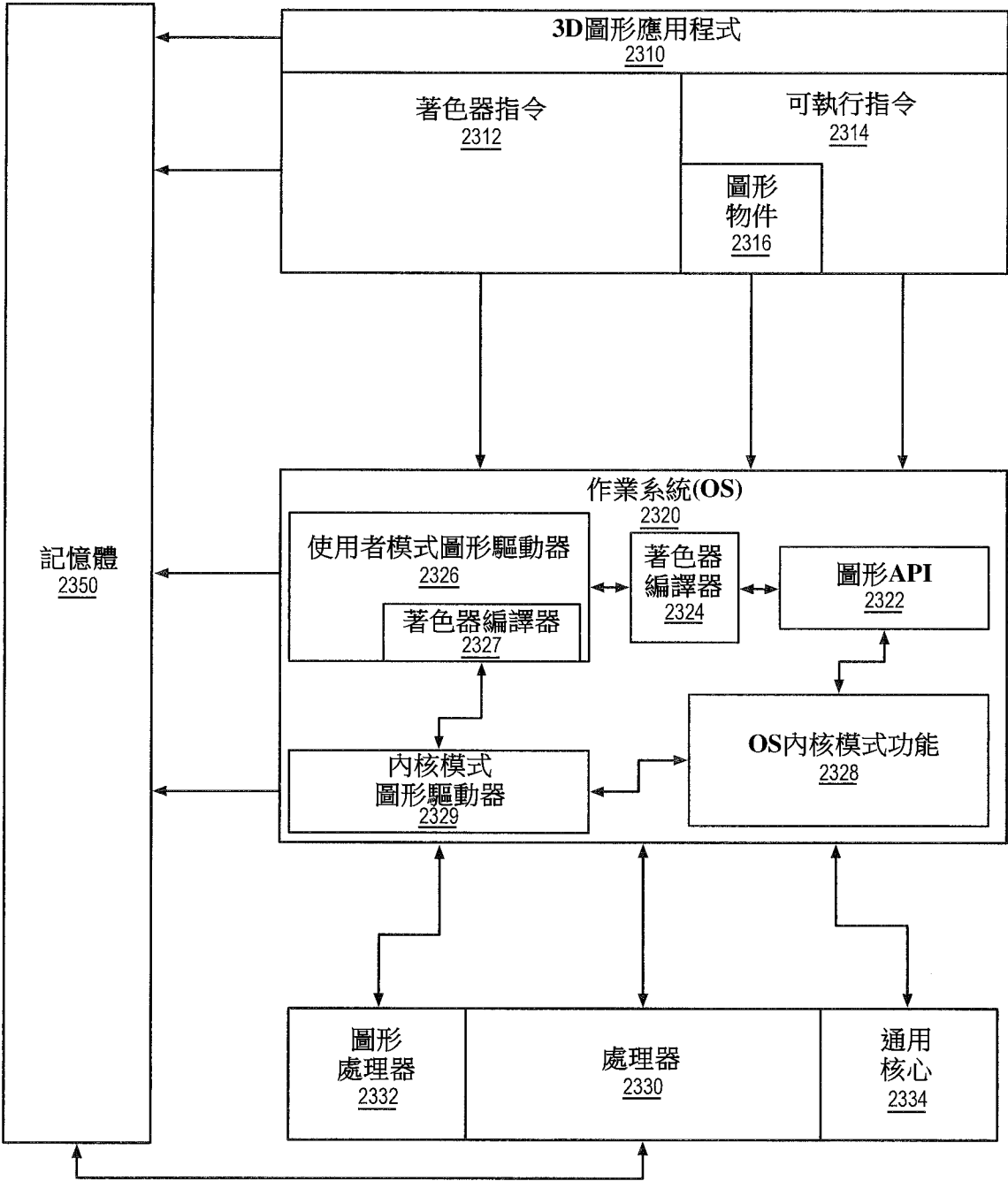
【圖 22A】

圖形處理器命令序列
2210



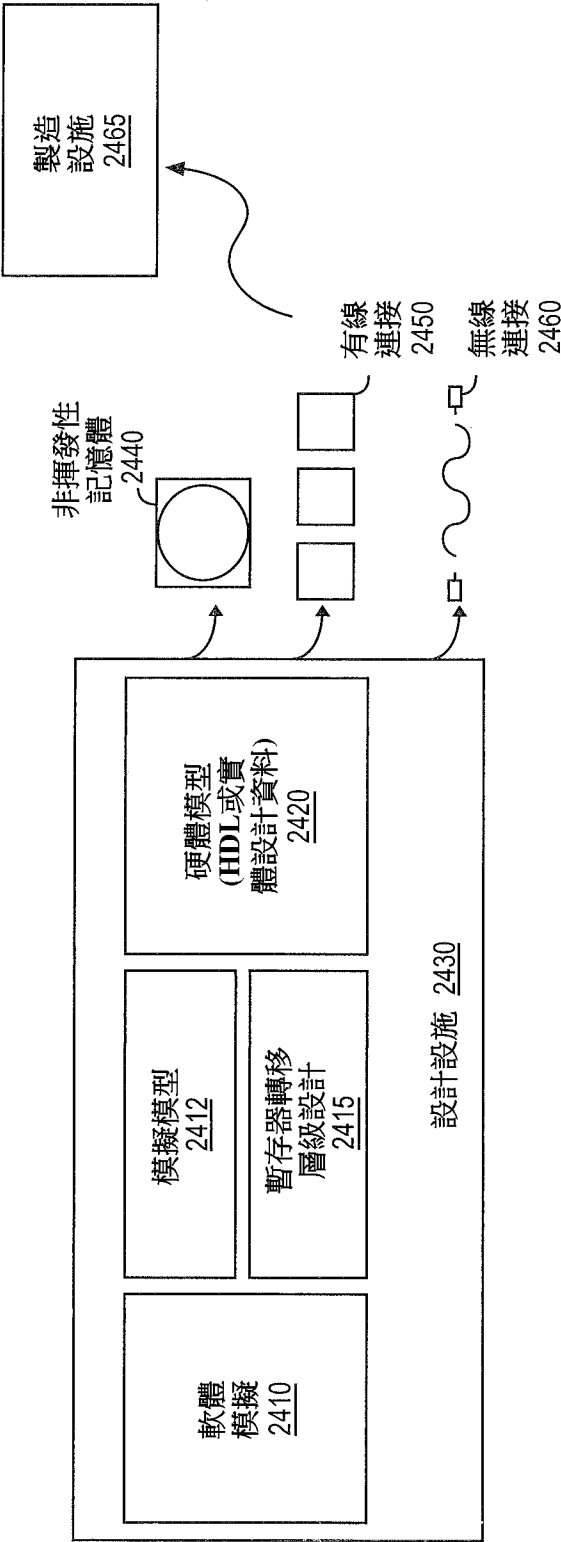
【圖 22B】

資料處理系統 - 2300

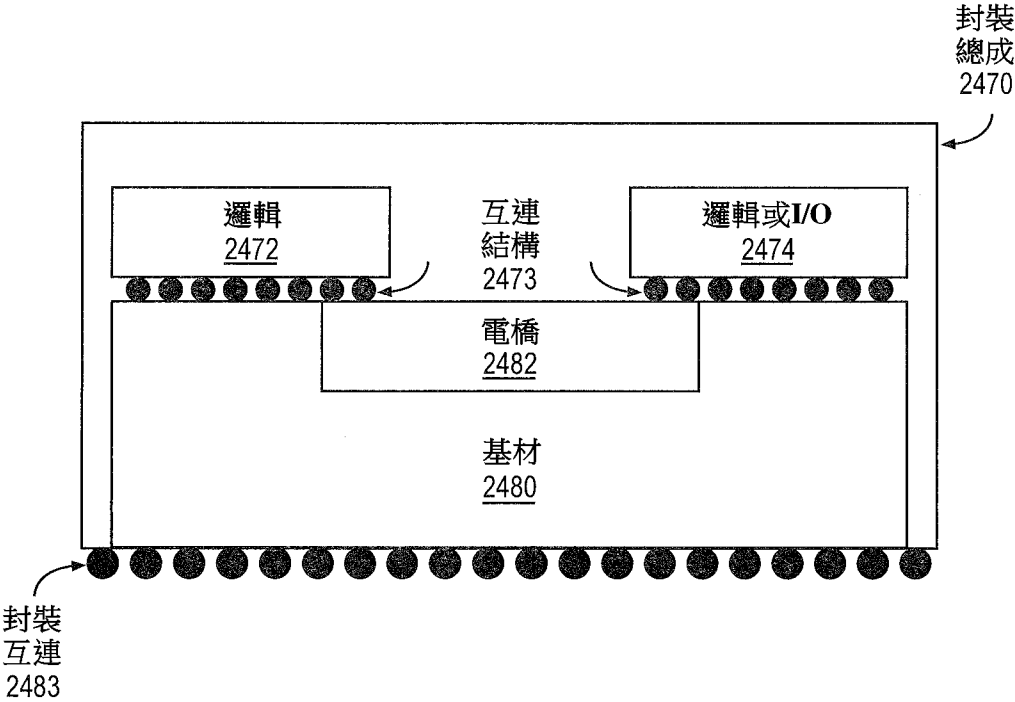


【圖 23】

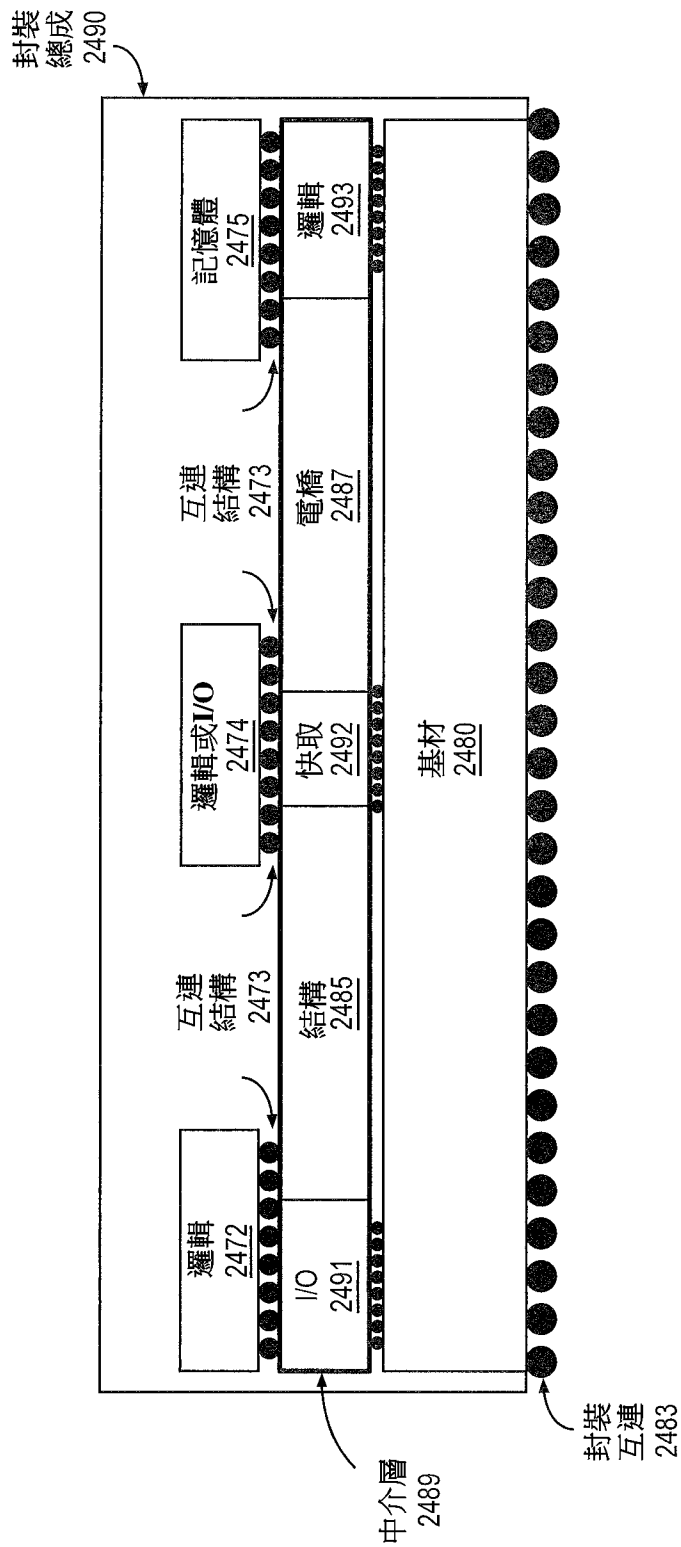
IP核心開發 - 2400



【圖 24A】

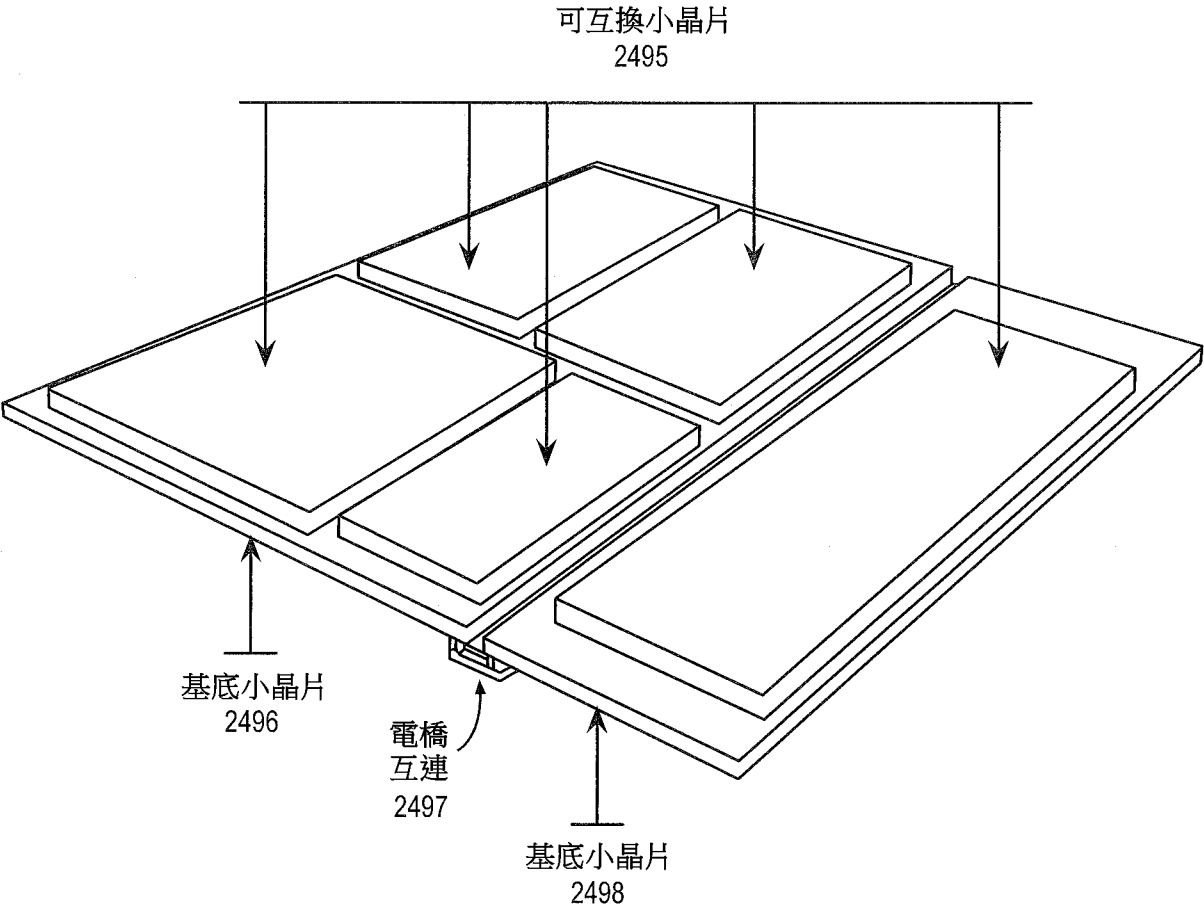


【圖 24B】

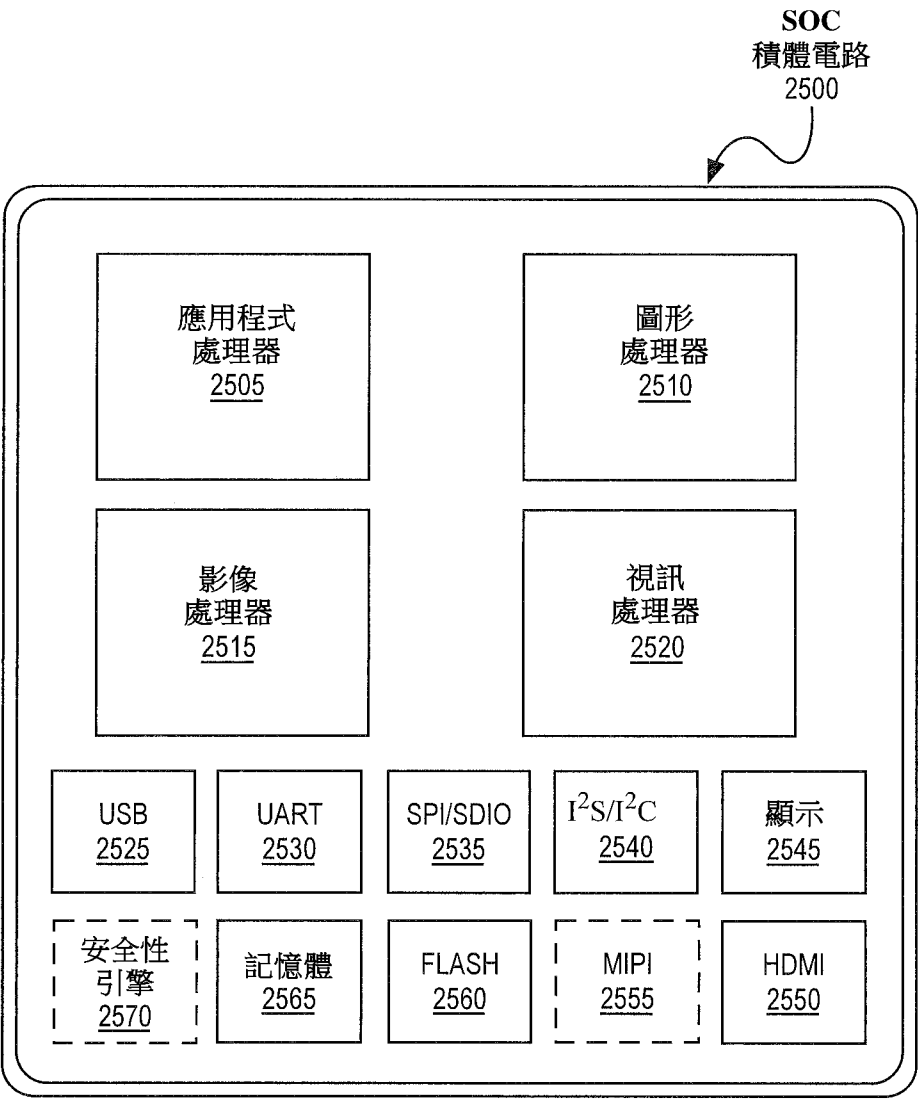


【圖 24C】

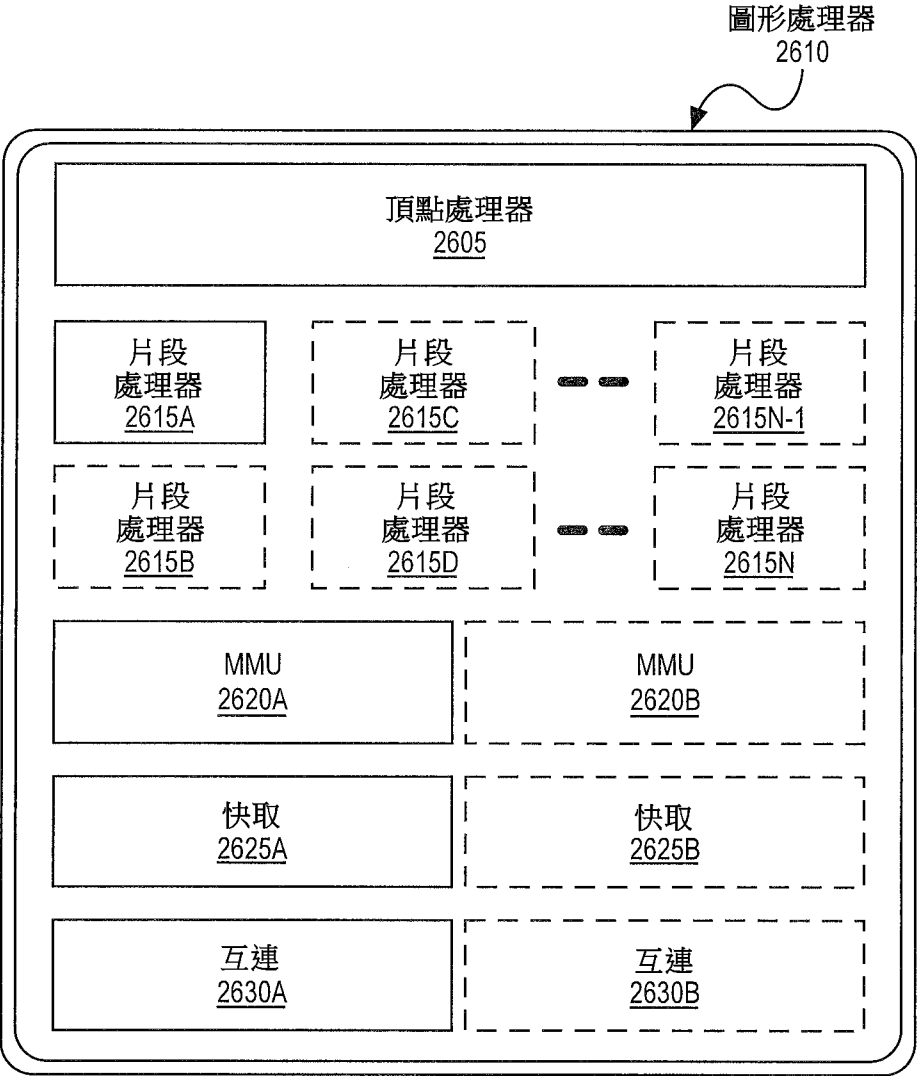
2494



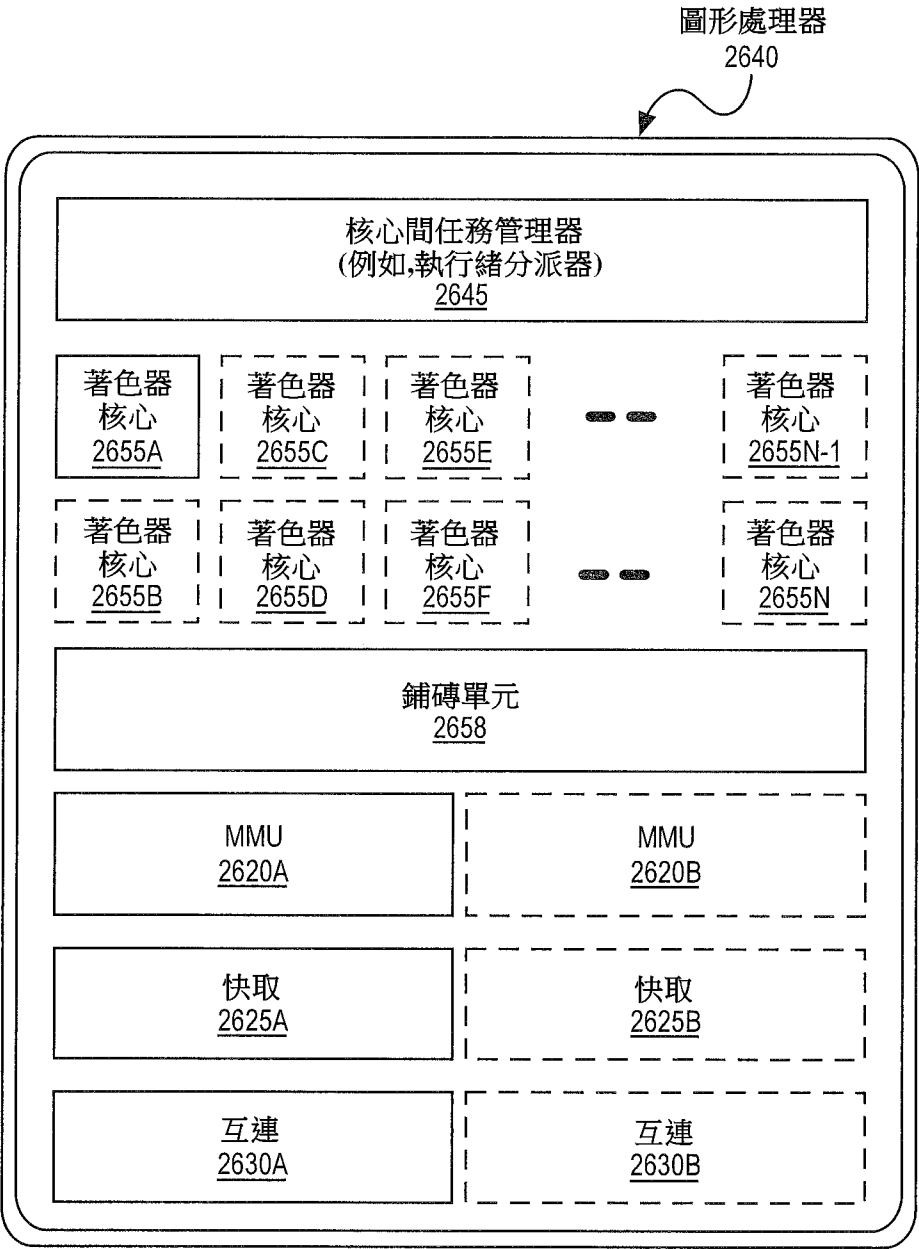
【圖 24D】



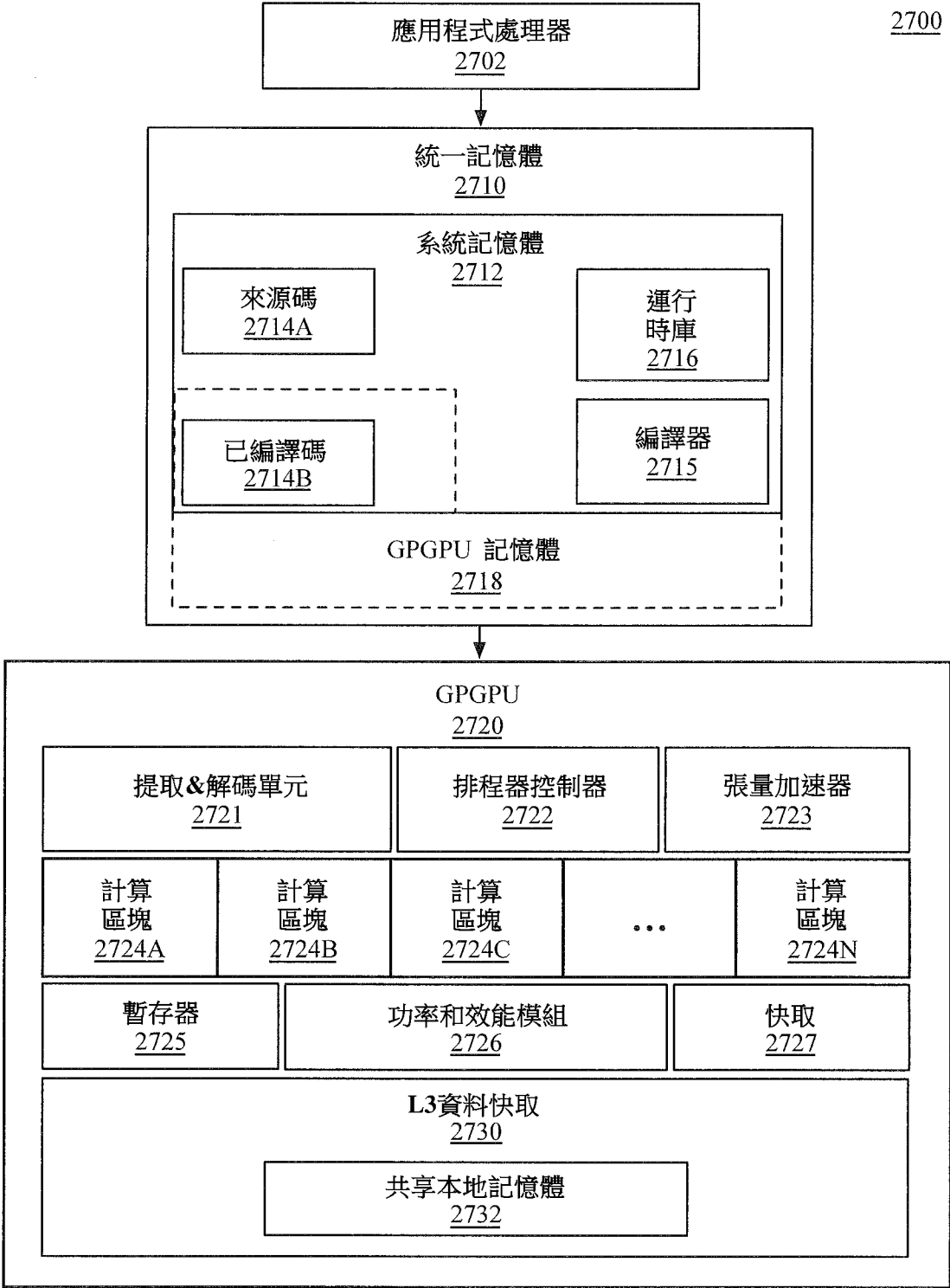
【圖 25】



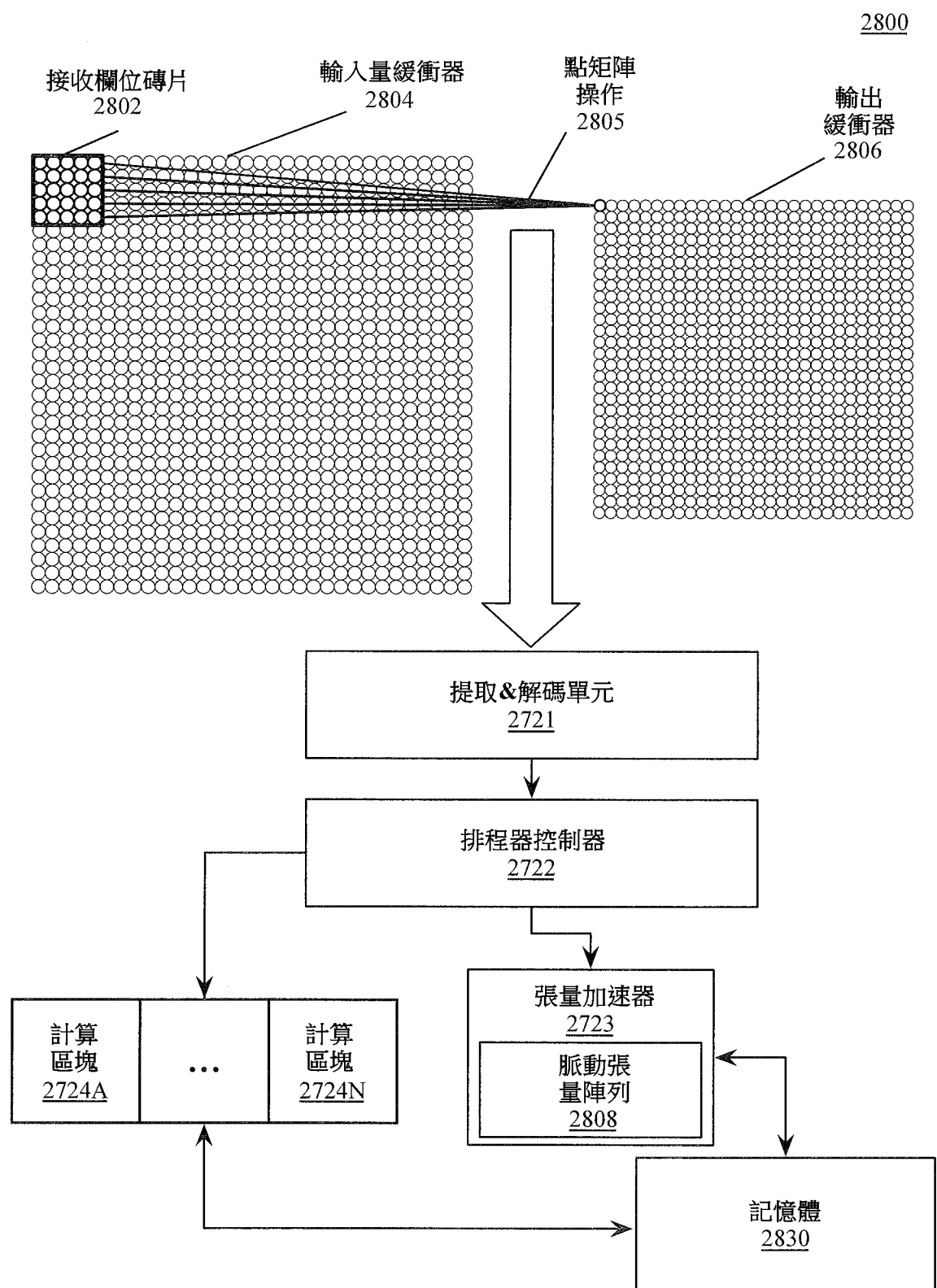
【圖 26A】



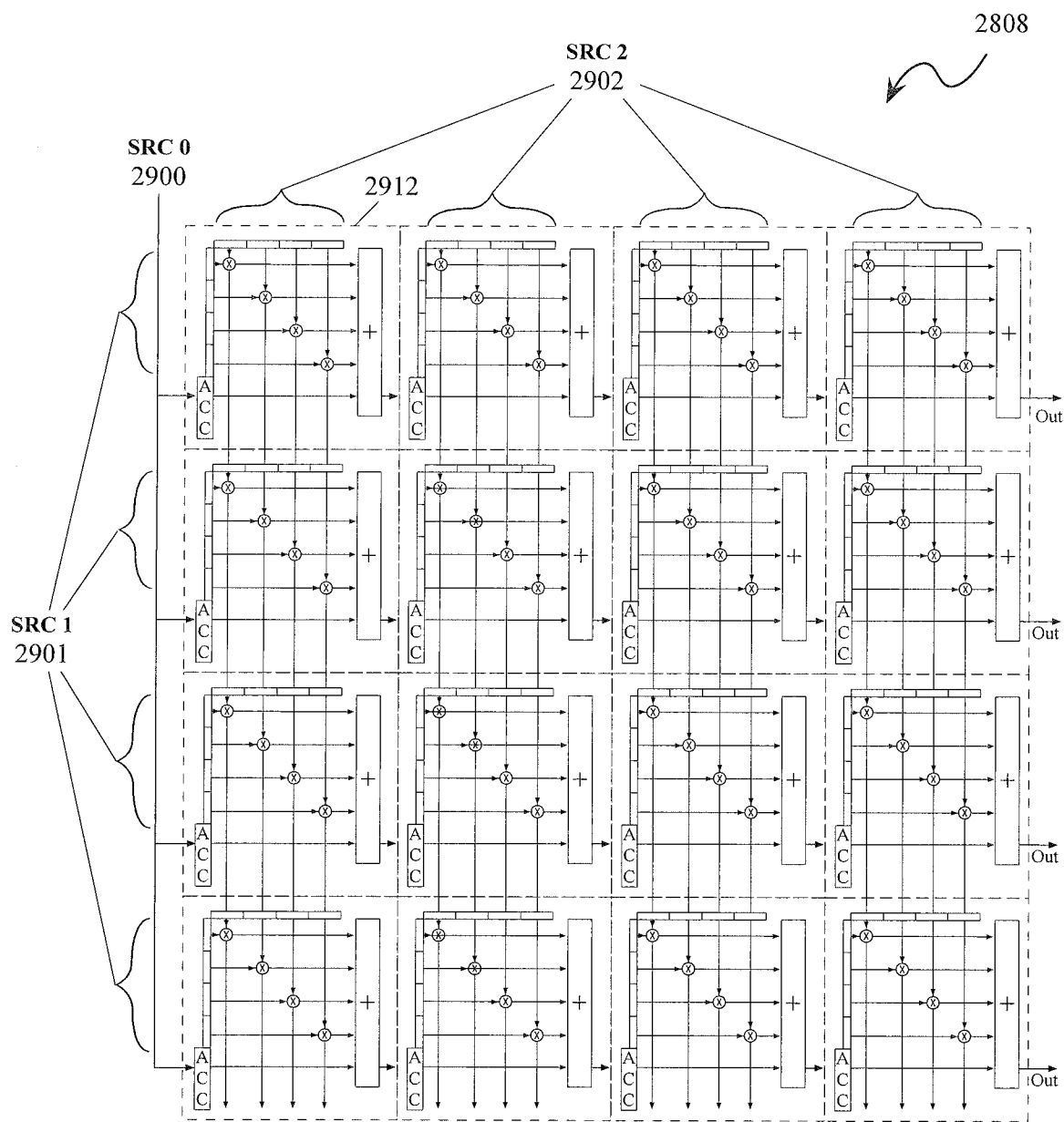
【圖 26B】



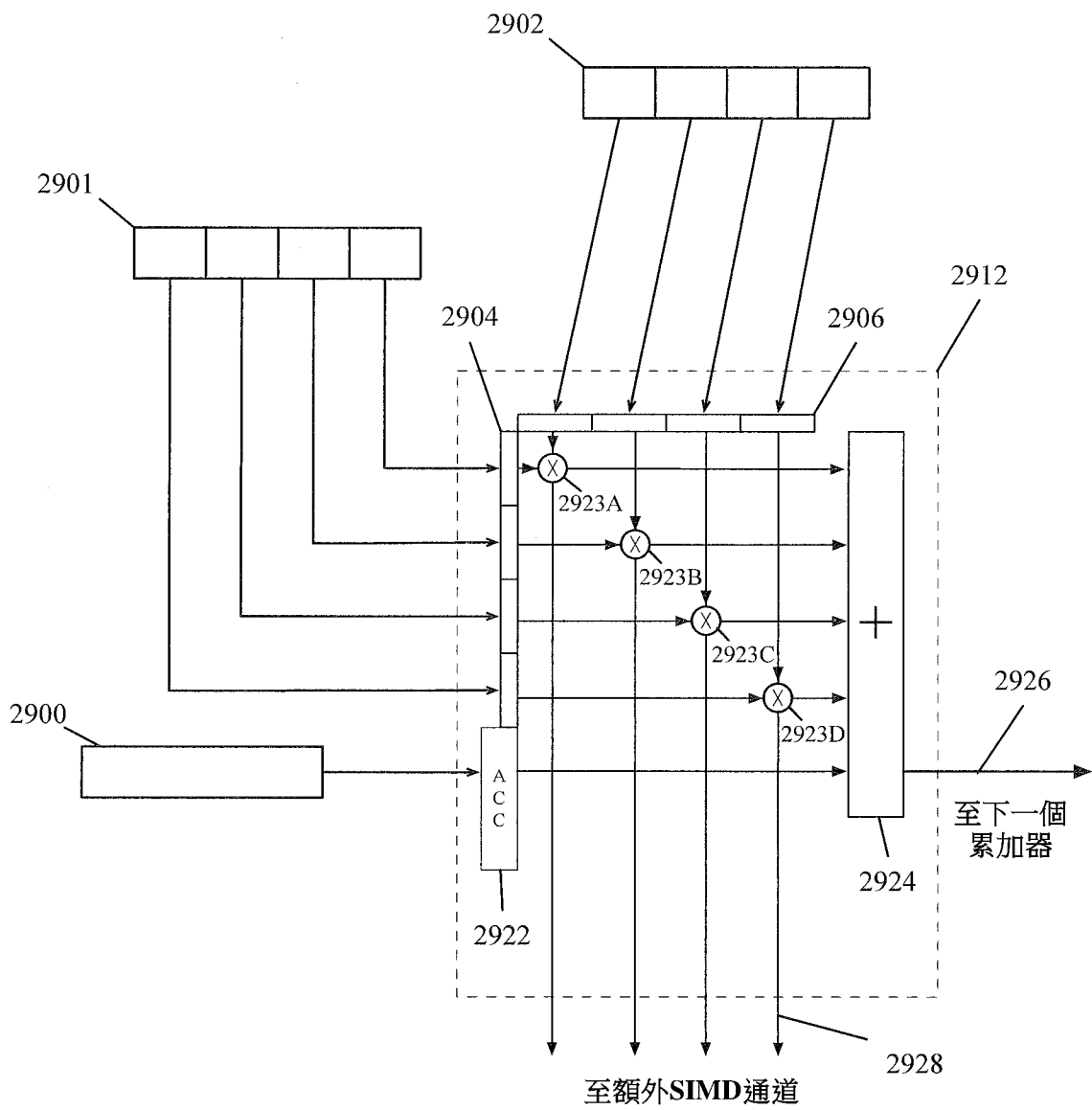
【圖 27】



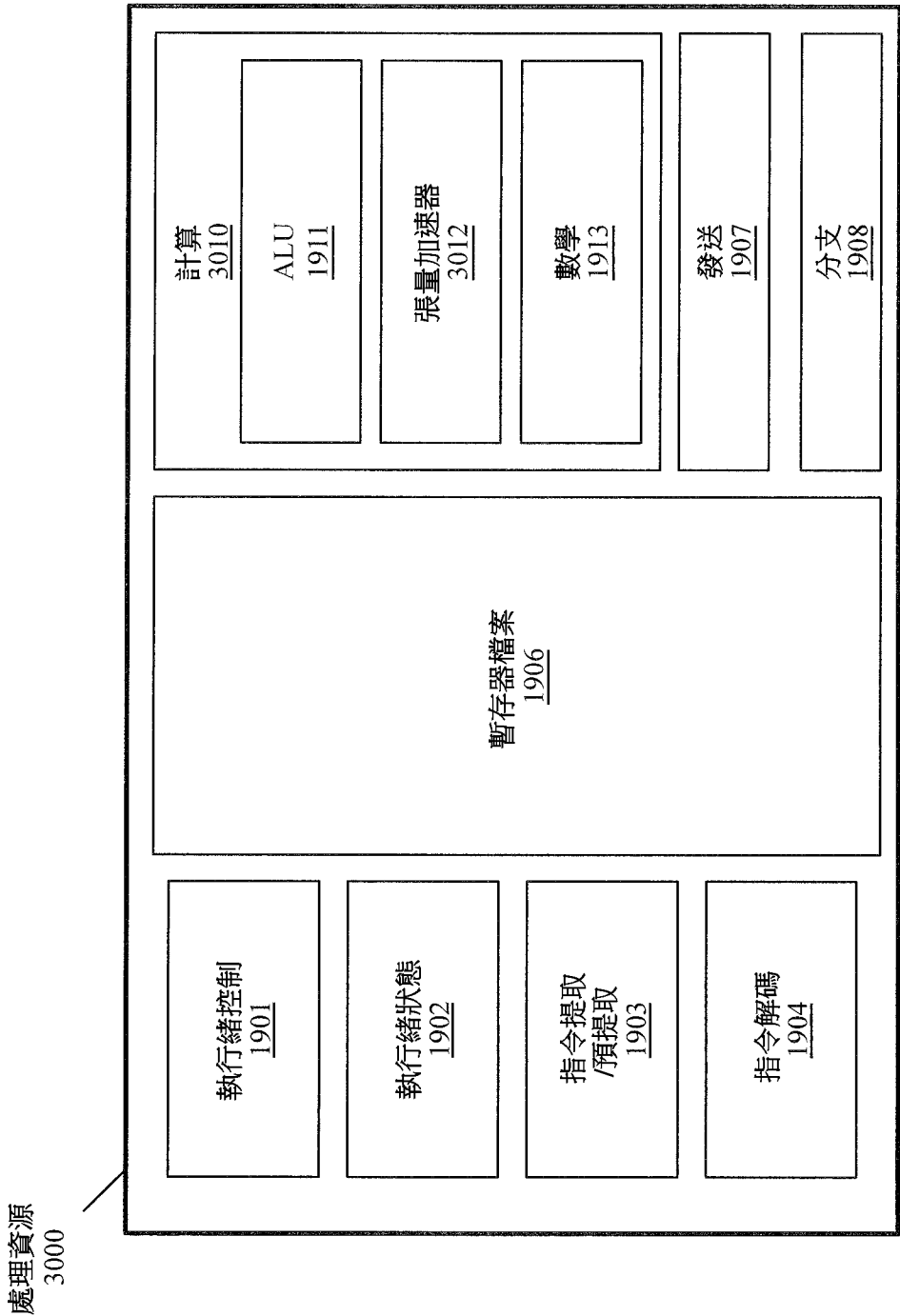
【圖 28】



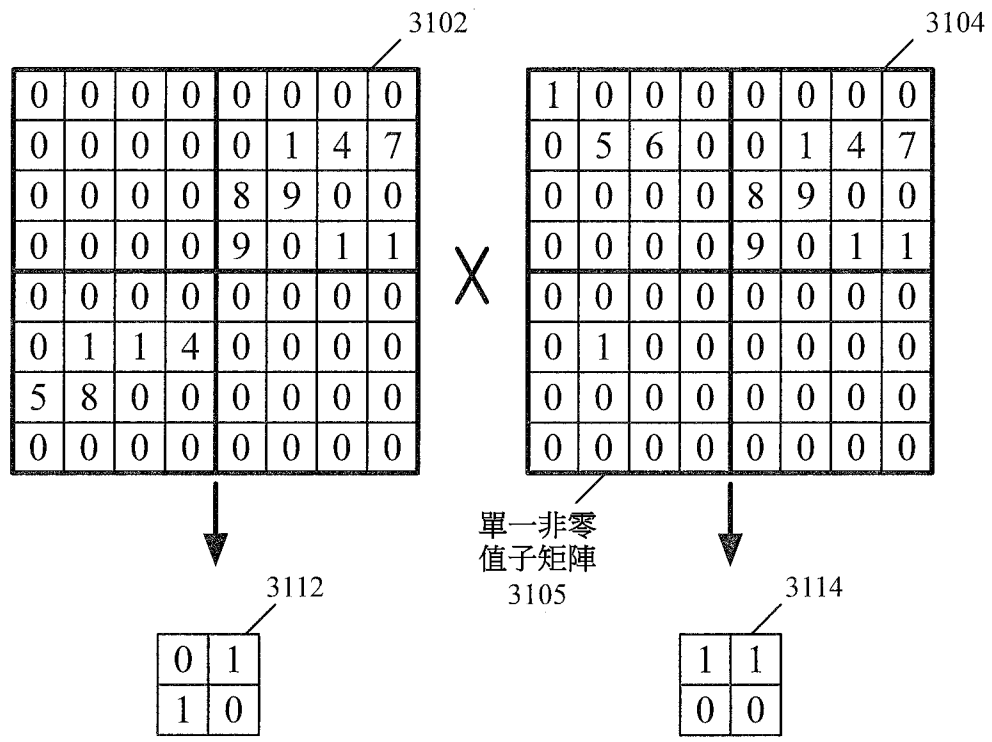
【圖 29A】



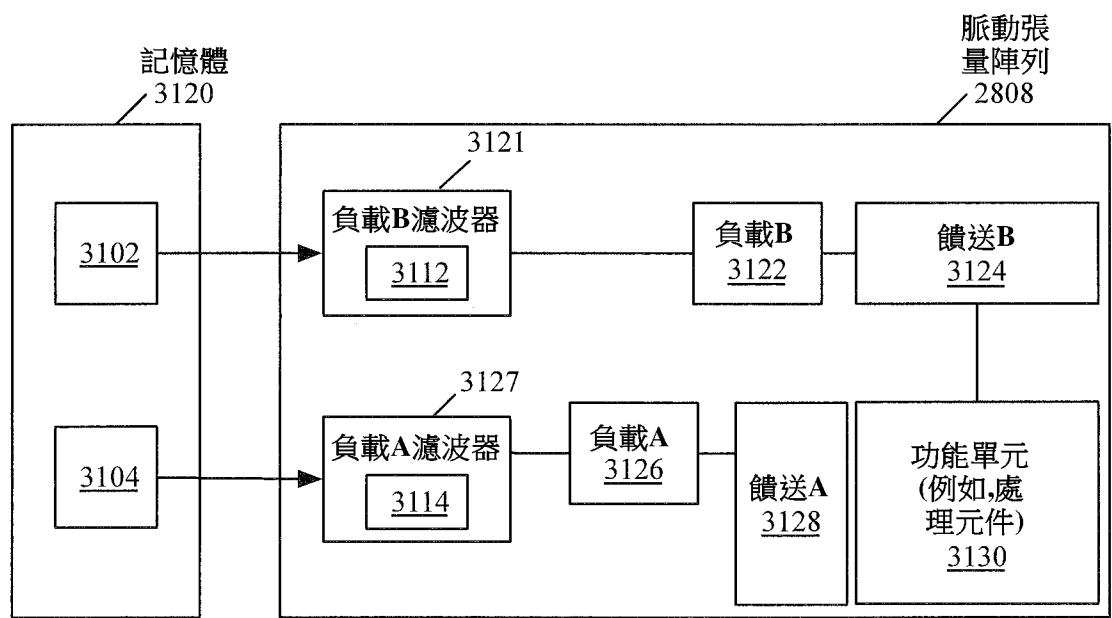
【圖 29B】



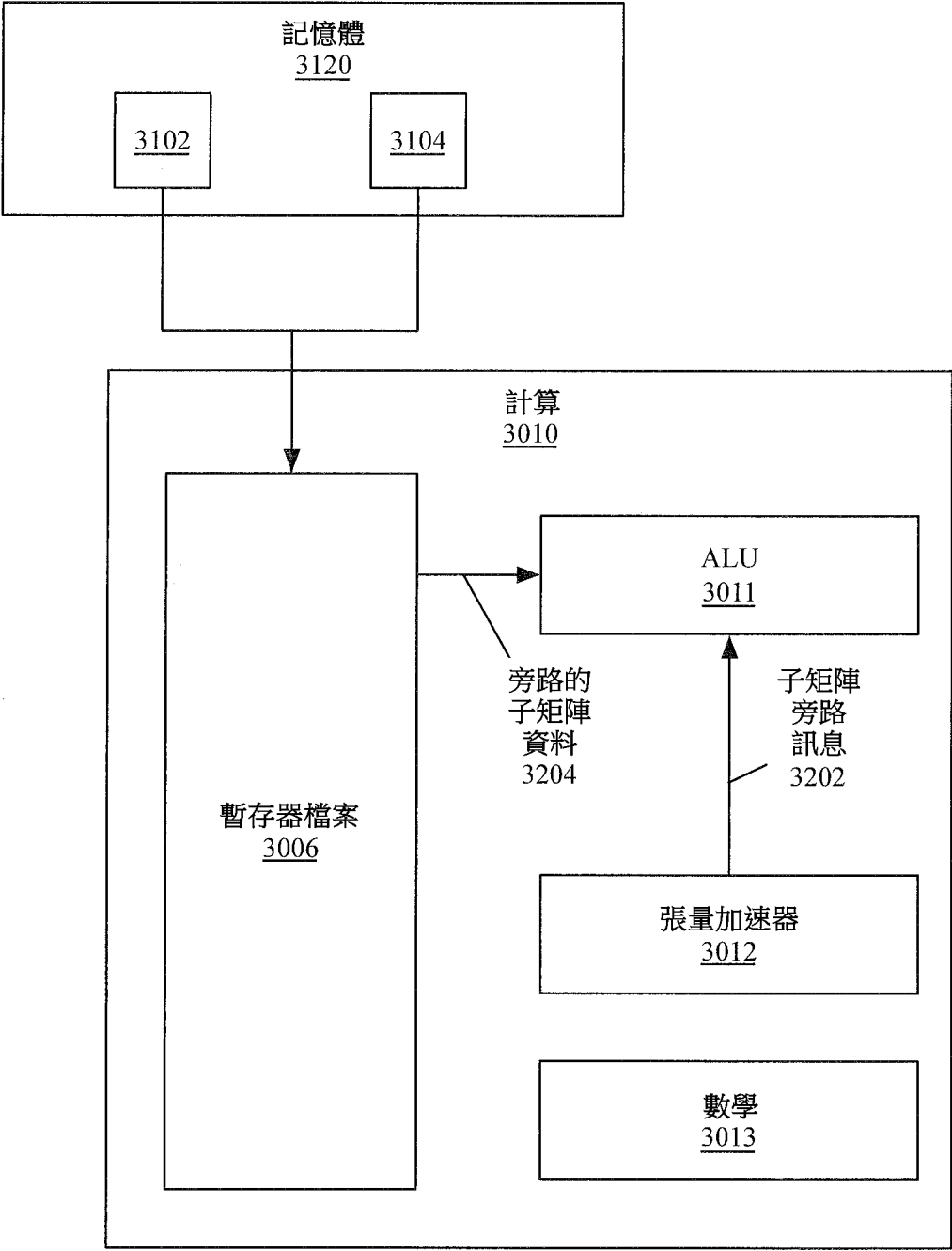
【圖 30】



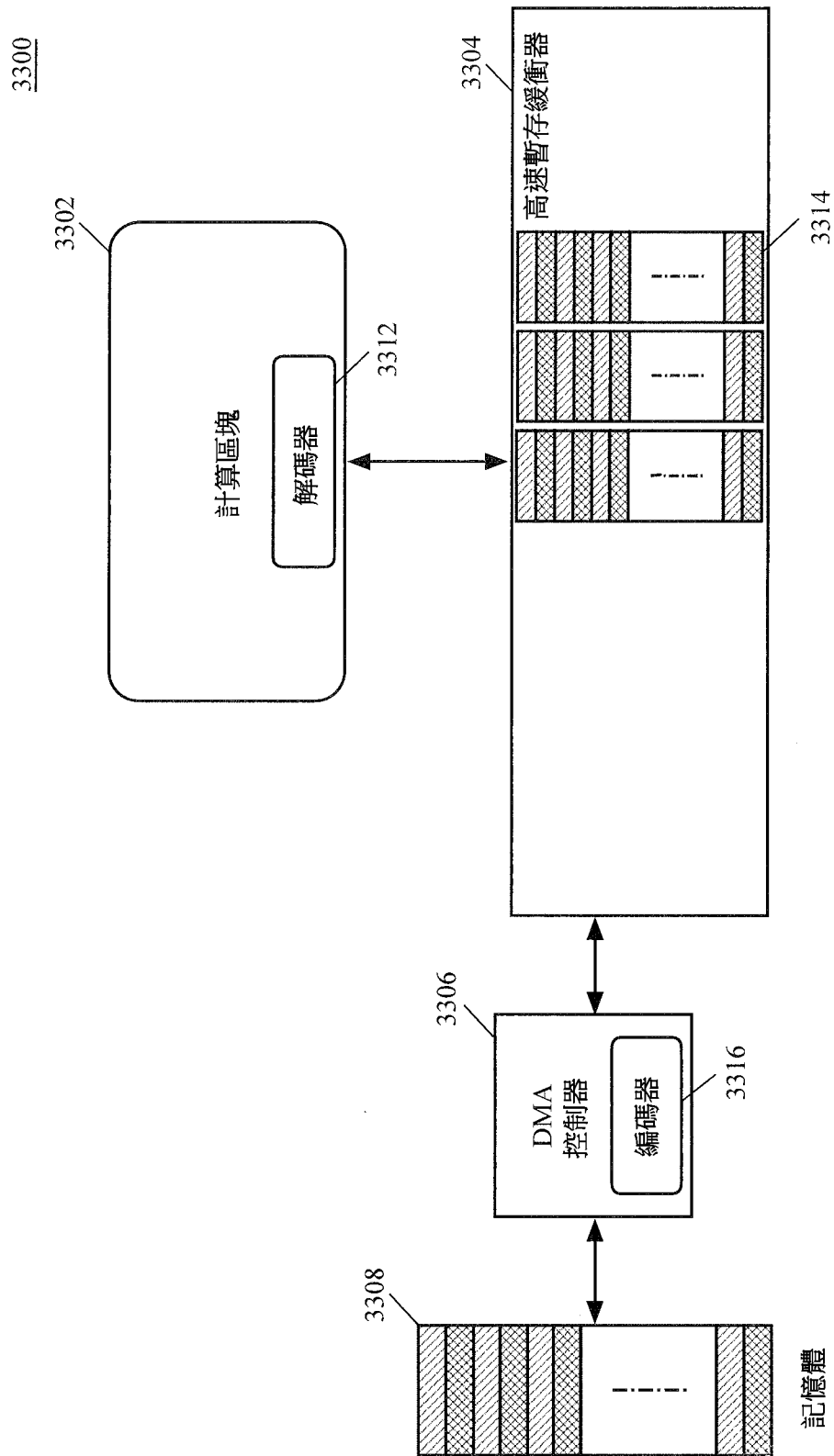
【圖 31A】



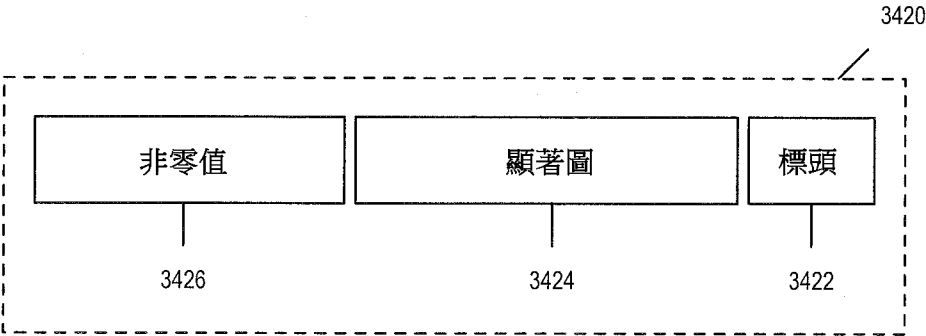
【圖 31B】



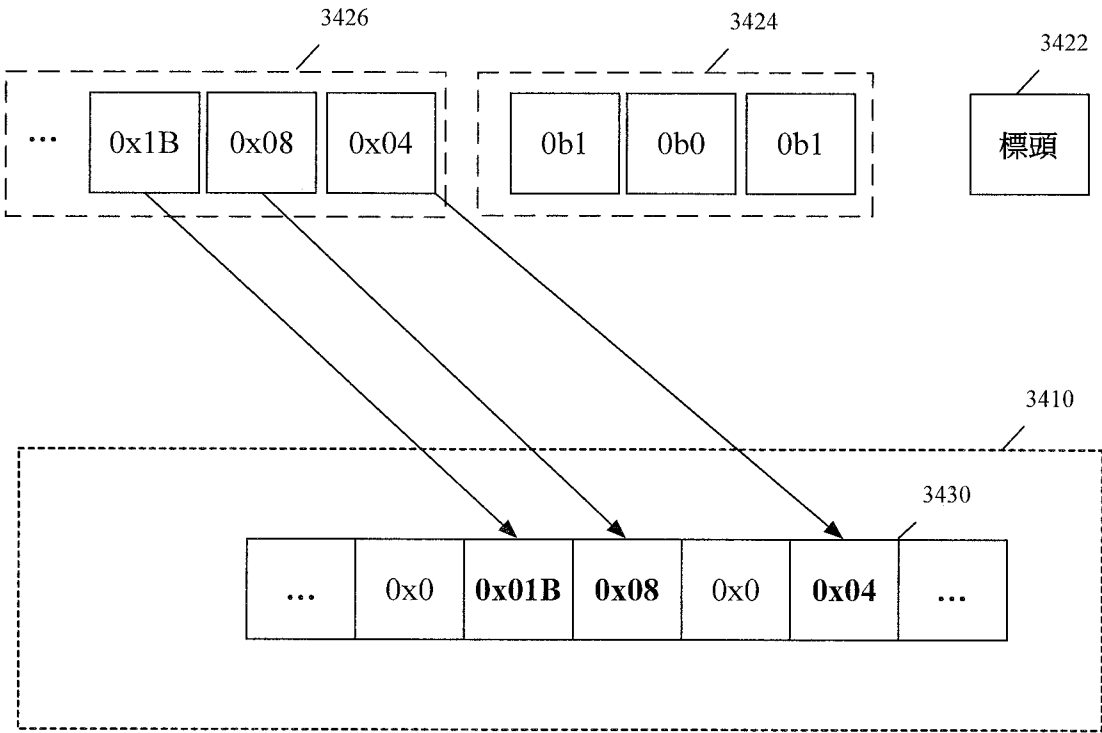
【圖 32】



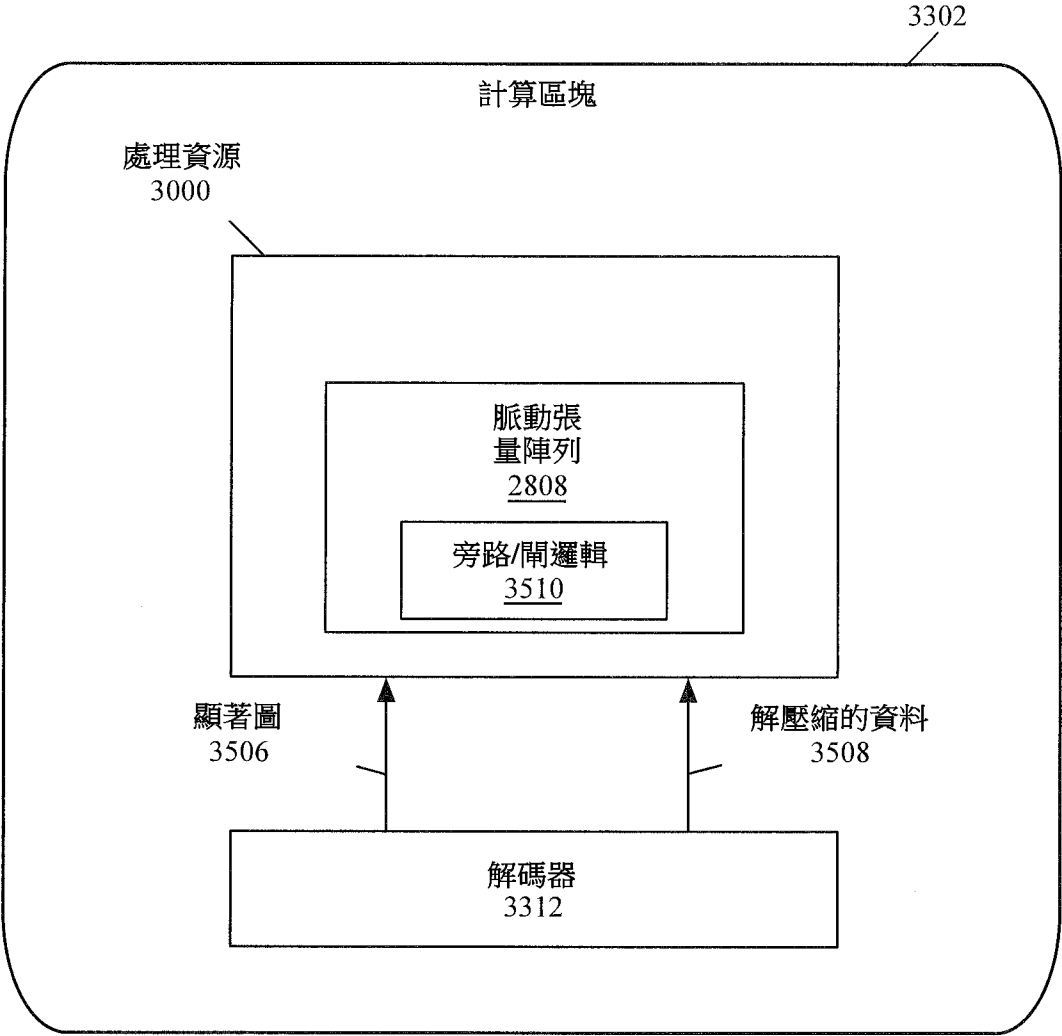
【圖 33】



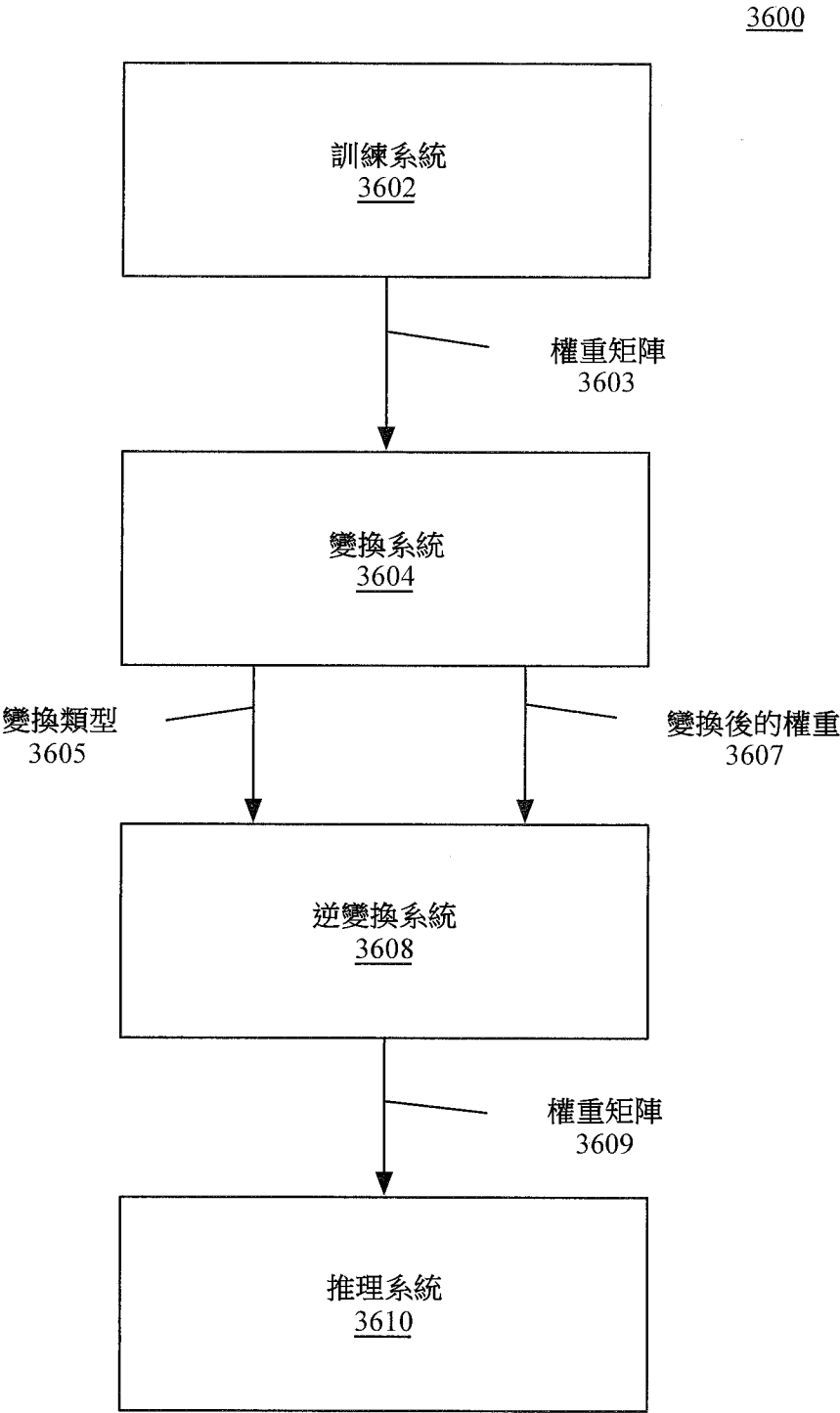
【圖 34A】



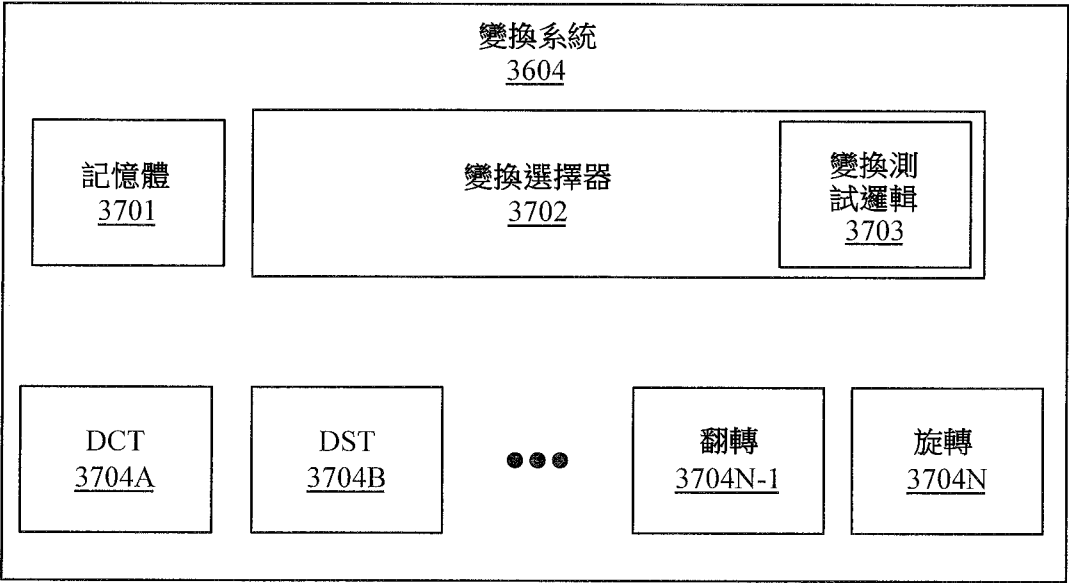
【圖 34B】



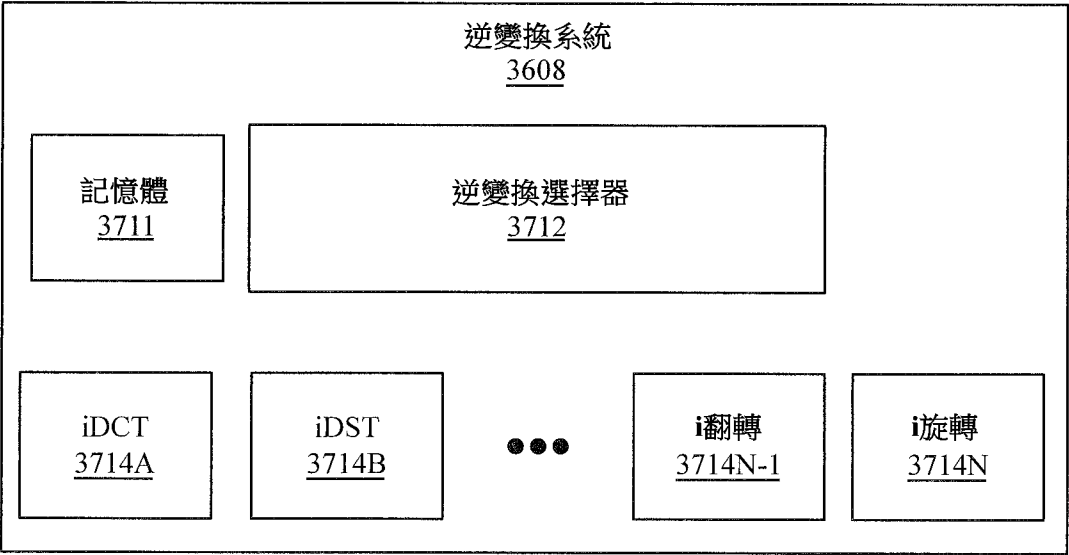
【圖 35】



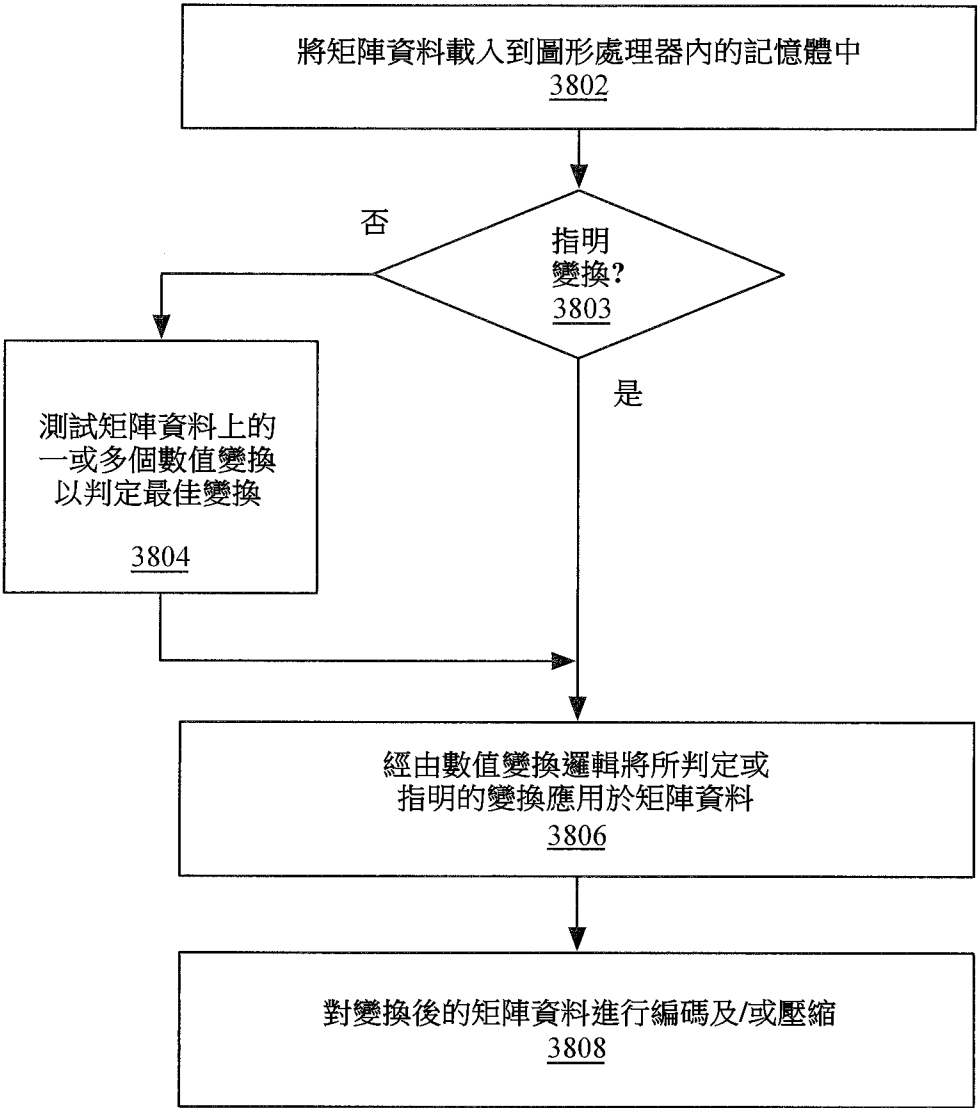
【圖 36】



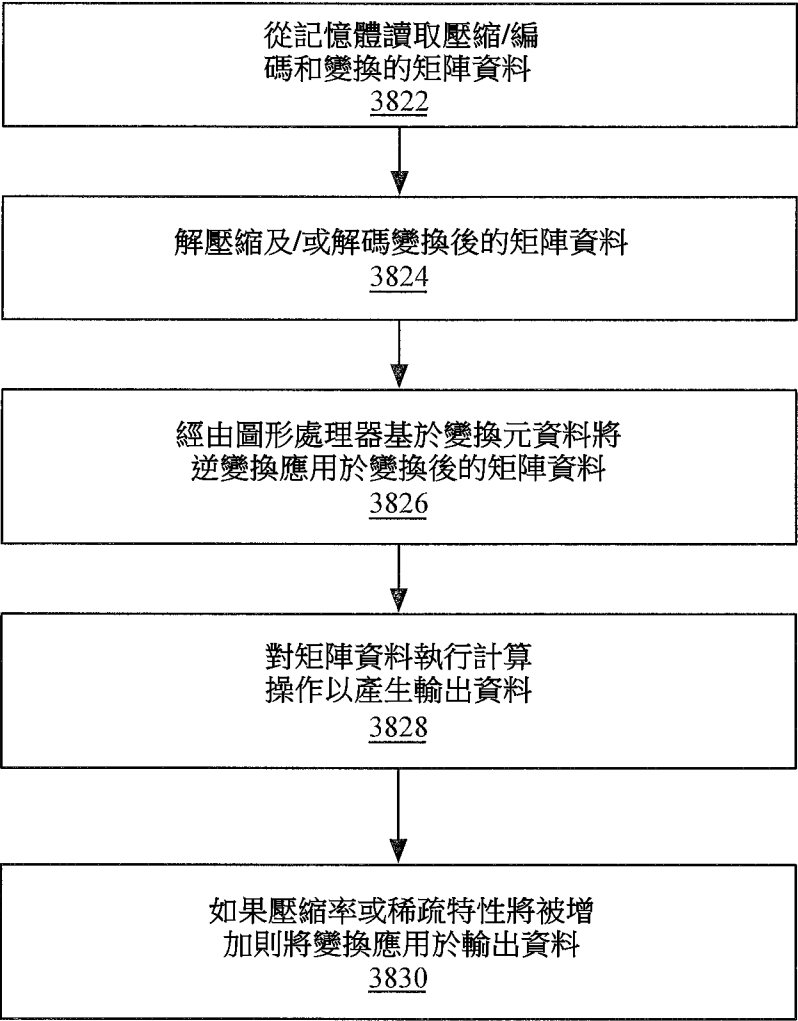
【圖 37A】



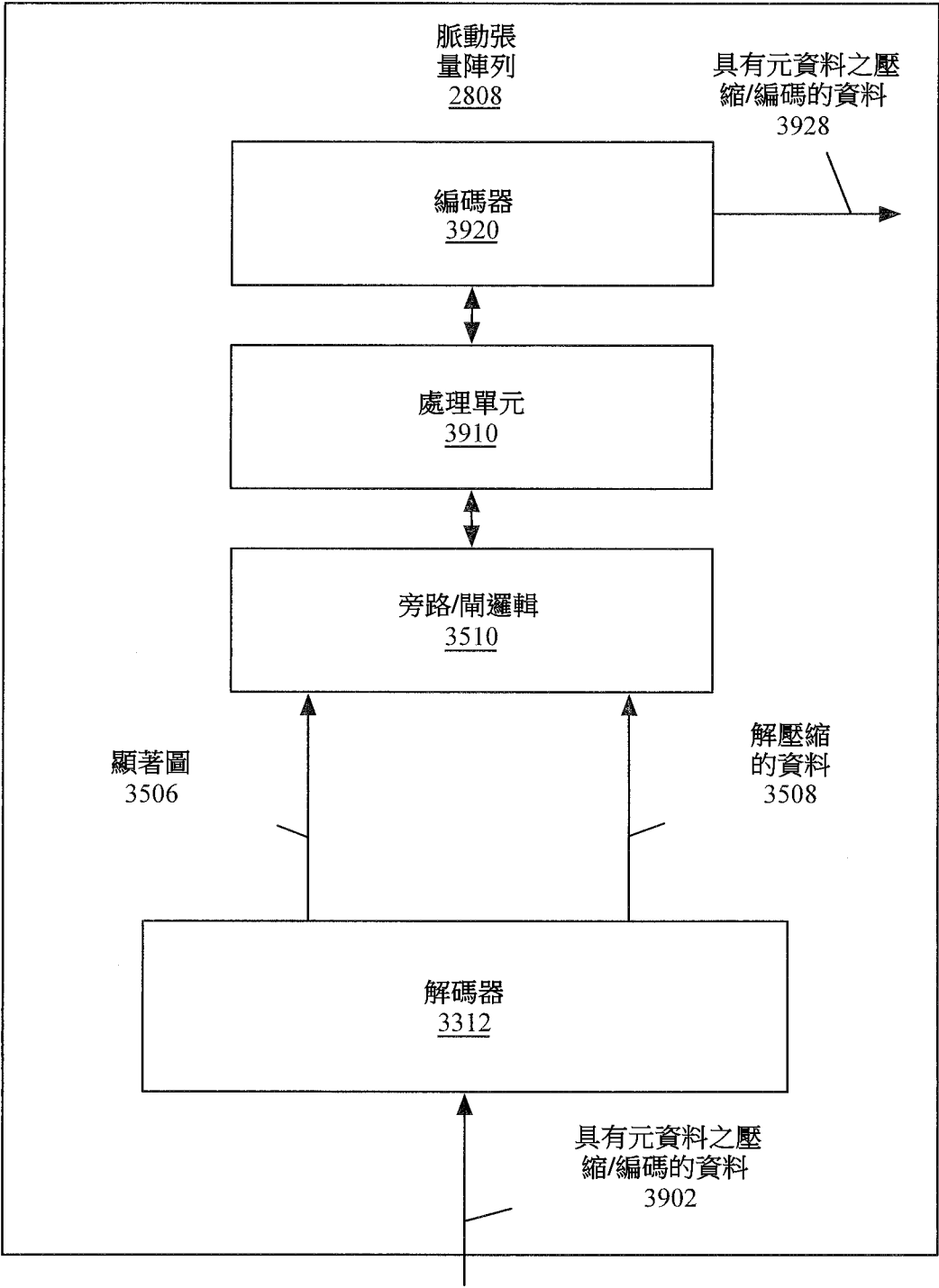
【圖 37B】



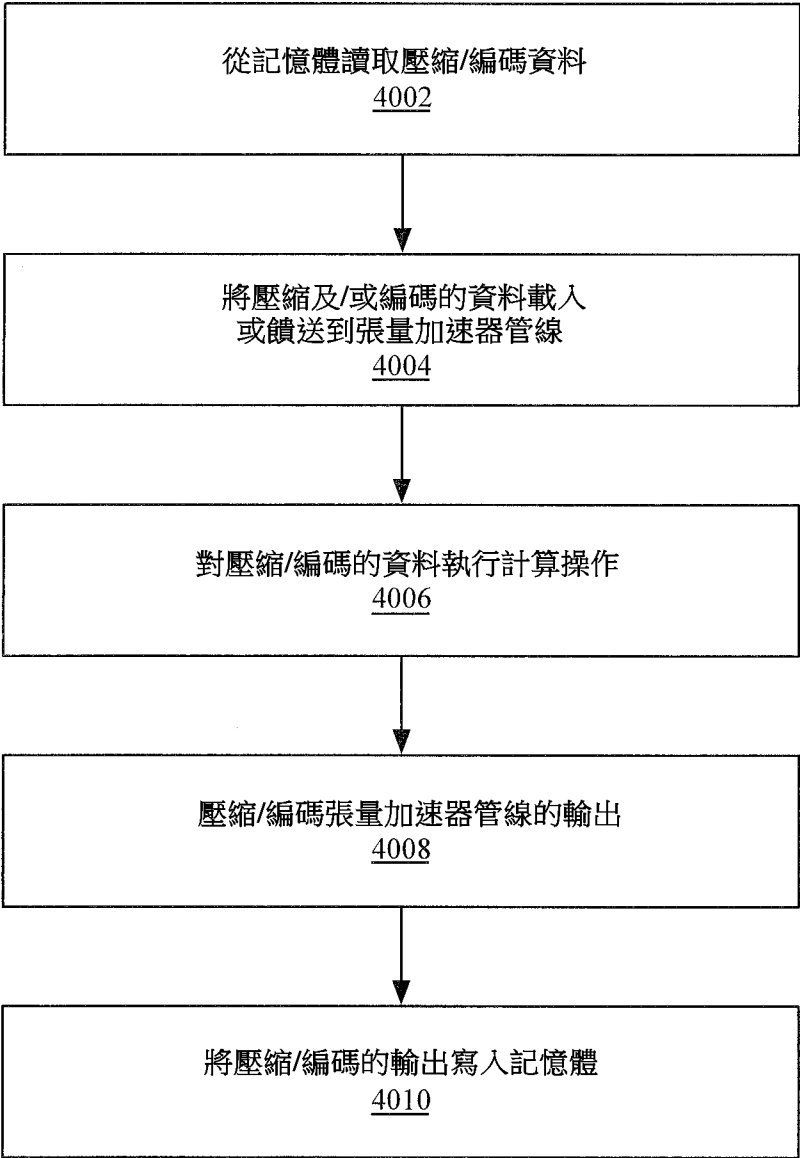
【圖 38A】



【圖 38B】



【圖 39】



【圖 40】



【圖 41】