



(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 I881126 B

(45)公告日：中華民國 114 (2025) 年 04 月 21 日

(21)申請案號：110118438

(22)申請日：中華民國 110 (2021) 年 05 月 21 日

(51)Int. Cl. : G06F13/16 (2006.01)

G06F3/06 (2006.01)

G06N20/00 (2019.01)

(30)優先權：2020/05/26 美國

16/883,869

(71)申請人：美商高通公司 (美國) QUALCOMM INCORPORATED (US)

美國

(72)發明人：韓德森 喬布 HENDERSON, JOAB (US)；坎貝爾 丹尼爾 CAMPBELL, DANIEL (US)；潘尼頓 萊恩 PENNINGTON, RYAN (US)

(74)代理人：李世章

(56)參考文獻：

US 2013/0117761A1

US 2013/0305009A1

US 2019/0115063A1

WO 2009/149236A1

審查人員：陳延慶

申請專利範圍項數：30 項 圖式數：10 共 65 頁

(54)名稱

用於記憶體中的推理的技術的增強型記憶體模組、方法及非暫時性電腦可讀取媒體

(57)摘要

本案的某些態樣提供了一種用於利用包括計算核心的增強型記憶體模組處理資料的方法，包括以下步驟：在增強型記憶體模組處從主機處理系統接收資料；將資料儲存在主機處理系統可定址記憶體中；從主機處理系統可定址記憶體向計算核心可定址記憶體傳遞資料；利用增強型記憶體模組上的計算核心處理資料以產生經處理的資料；從計算核心可定址記憶體向主機處理系統可定址記憶體傳遞經處理的資料；及經由主機處理系統可定址記憶體向主機處理系統提供經處理的資料。

Certain aspects of the present disclosure provide a method for processing data with an enhanced memory module comprising a compute core, including: receiving data at the enhanced memory module from a host processing system; storing the data in host processing system-addressable memory; transferring the data from the host processing system-addressable memory to compute core-addressable memory; processing the data with the compute core on the enhanced memory module to generate processed data; transferring the processed data from the compute core-addressable memory to the host processing system-addressable memory; and providing the processed data to the host processing system via the host processing system-addressable memory.

指定代表圖：

符號簡單說明：

- 200:處理系統
- 202:OS 使用者空間
- 204:模型
- 206:資料
- 208:ML/AI 應用軟體
- 210:加速器運行時間應用軟體
- 212:OS 核心空間
- 214:記憶體驅動器
- 216:加速器驅動器
- 218:CPU
- 220:記憶體管理單元 (MMU)
- 222:記憶體控制器
- 224A:增強型記憶體模組
- 224B:增強型記憶體模組
- 226A:驅動器
- 226B:驅動器

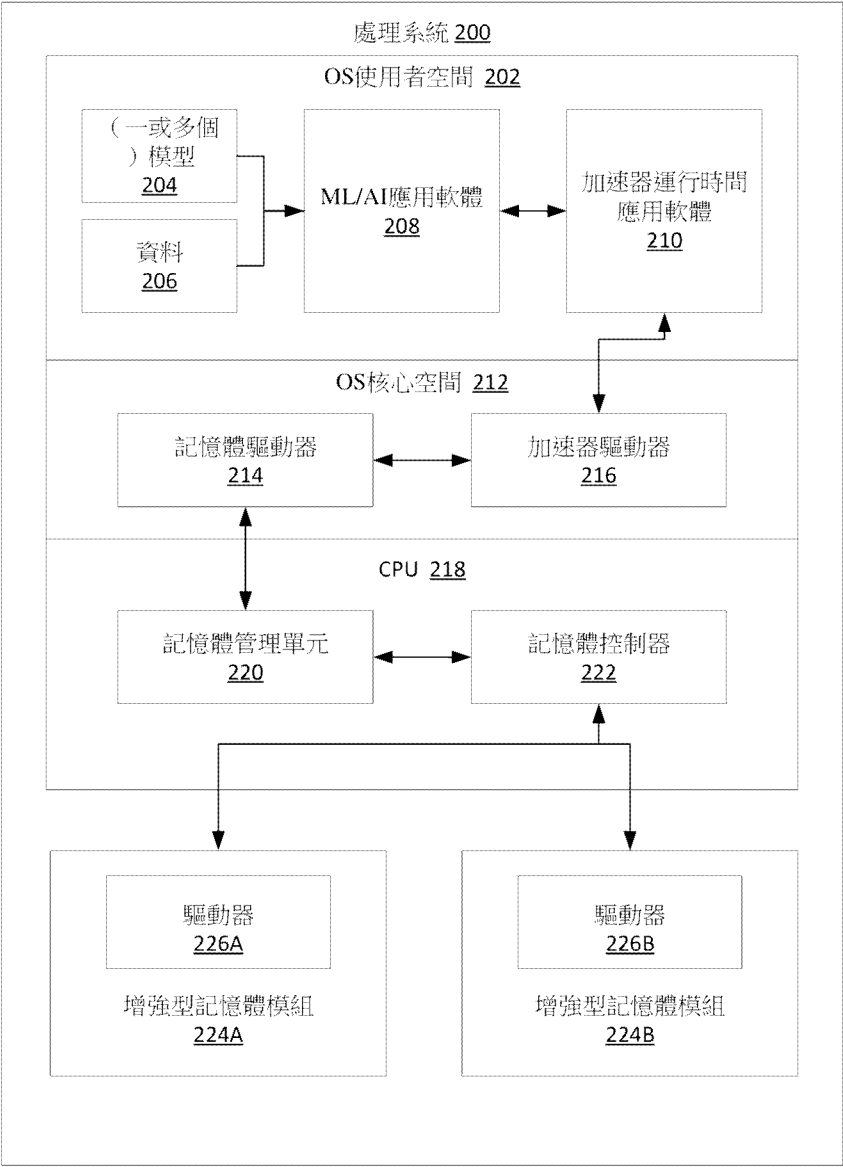


圖2



I881126

【發明摘要】

【中文發明名稱】用於記憶體中的推理的技術的增強型記憶體模組、方法及非暫時性電腦可讀取媒體

【英文發明名稱】 ENHANCED MEMORY MODULE, METHOD, AND NON-TRANSITORY COMPUTER-READABLE MEDIUM FOR TECHNIQUES FOR INFERENCE IN MEMORY

【中文】

本案的某些態樣提供了一種用於利用包括計算核心的增強型記憶體模組處理資料的方法，包括以下步驟：在增強型記憶體模組處從主機處理系統接收資料；將資料儲存在主機處理系統可定址記憶體中；從主機處理系統可定址記憶體向計算核心可定址記憶體傳遞資料；利用增強型記憶體模組上的計算核心處理資料以產生經處理的資料；從計算核心可定址記憶體向主機處理系統可定址記憶體傳遞經處理的資料；及經由主機處理系統可定址記憶體向主機處理系統提供經處理的資料。

【英文】

Certain aspects of the present disclosure provide a method for processing data with an enhanced memory module comprising a compute core, including: receiving data at the enhanced memory module from a host processing system; storing the data in host processing system-addressable memory; transferring the data from the host processing system-addressable memory to compute core-addressable memory; processing the data with the compute core on the enhanced memory module to generate processed data; transferring the processed data from the compute core-addressable memory to the

host processing system-addressable memory; and providing the processed data to the host processing system via the host processing system-addressable memory.

【指定代表圖】第（ 2 ）圖。

【代表圖之符號簡單說明】

2 0 0 : 處 理 系 統

2 0 2 : O S 使 用 者 空 間

2 0 4 : 模 型

2 0 6 : 資 料

2 0 8 : M L / A I 應 用 軟 體

2 1 0 : 加 速 器 運 行 時 間 應 用 軟 體

2 1 2 : O S 核 心 空 間

2 1 4 : 記 憶 體 驅 動 器

2 1 6 : 加 速 器 驅 動 器

2 1 8 : C P U

2 2 0 : 記 憶 體 管 理 單 元 (M M U)

2 2 2 : 記 憶 體 控 制 器

2 2 4 A : 增 強 型 記 憶 體 模 組

2 2 4 B : 增 強 型 記 憶 體 模 組

2 2 6 A : 驅 動 器

2 2 6 B : 驅 動 器

【特徵化學式】

無

【發明說明書】

【中文發明名稱】用於記憶體中的推理的技術的增強型記憶體模組、方法及非暫時性電腦可讀取媒體

【英文發明名稱】 ENHANCED MEMORY MODULE, METHOD, AND NON-TRANSITORY COMPUTER-READABLE MEDIUM FOR TECHNIQUES FOR INFERENCE IN MEMORY

【技術領域】

【0001】 本案的態樣係關於在非傳統計算硬體中執行機器學習和人工智慧任務，並且具體地係關於在記憶體硬體中執行此類任務。

【先前技術】

【0002】 在大資料時代，對機器學習和人工智慧能力的需求急劇增加。照慣例，機器學習已被用於產生模型，該等模型隨後可以產生用於人工智慧任務的推理，諸如預測、分類等。隨著對推理能力需求的增加，計算硬體製造商正在尋求擴展現有計算平臺（諸如桌上型電腦和伺服器）中以及其他新興種類的處理系統（諸如行動設備和邊緣處理設備）中的推理能力的密度（density）。

【0003】 傳統上，機器學習和人工智慧「加速器」已經經由硬體擴展介面（例如，主機板上的PCIe插槽）而被添加到系統中，以擴展底層計算基礎結構的能力。不幸的是，將此類硬體擴展介面用於加速器意味著彼等相同的介面不能被用於其他目的，諸如聯網、圖形渲染、安全性、聲音處理和其他常見計算任務。因此，向現有處理系統添加附

加的機器學習和人工智慧最佳化計算硬體常常以犧牲其他必要的計算元件為代價。

【0004】 相應地，需要的是在不犧牲其他必要元件的情況下向現有處理系統添加附加的機器學習和人工智慧計算能力的系統和方法。

【發明內容】

【0005】 本案的某些態樣提供了一種增強型記憶體模組，包括：計算核心；及一或多個雙模式記憶體元件，其中增強型記憶體模組被配置為：將一或多個雙模式記憶體元件中的記憶體的第一子集分配為主機處理系統可定址記憶體，並且將一或多個雙模式記憶體元件中的記憶體的第二子集分配為計算核心可定址記憶體；從主機處理系統接收資料；利用計算核心處理資料以產生經處理的資料；及經由記憶體的第一子集向主機處理系統提供經處理的資料。

【0006】 進一步的態樣提供了一種用於利用包括計算核心的增強型記憶體模組處理資料的方法，包括以下步驟：經由將增強型記憶體模組的記憶體的第一子集分配為主機處理系統可定址記憶體，並且將增強型記憶體模組的記憶體的第二子集分配為計算核心可定址記憶體來初始化增強型記憶體模組；在增強型記憶體模組處從主機處理系統接收資料；利用增強型記憶體模組上的計算核心處理資料以產生經處理的資料；及經由記憶體的第一子集向主機處理系統提供經處理的資料。

【0007】 進一步的態樣提供了一種包括指令的非暫時性電腦可讀取媒體，該等指令在被主機處理系統中的一或多個處理器執行時，使得處理系統執行用於利用包括計算核心的增強型記憶體模組處理資料的方法，該方法包括以下步驟：經由將增強型記憶體模組的記憶體的第一子集分配為主機處理系統可定址記憶體，並且將增強型記憶體模組的記憶體的第二子集分配為計算核心可定址記憶體來初始化增強型記憶體模組；在增強型記憶體模組處從主機處理系統接收資料；利用增強型記憶體模組上的計算核心處理資料以產生經處理的資料；及經由記憶體的第一子集向主機處理系統提供經處理的資料。

【0008】 下文的描述和相關附圖詳細說明一或多個實施例的某些說明性特徵。

【圖式簡單說明】

【0009】 所附附圖圖示了一或多個實施例的某些態樣，並且因此不應被認為是對本案的範疇的限制。

【0010】 圖 1 圖示了包括記憶體整合加速器 (memory-integrated-accelerator) 的增強型記憶體模組的實例。

【0011】 圖 2 圖示了包括如參考圖 1 所描述的增強型記憶體模組的示例性處理系統。

【0012】 圖 3 A - 圖 3 C 圖示了用於增強型記憶體模組的示例性配置。

【0013】 圖 4 圖示了用於向增強型記憶體模組上的計算核心記憶體元件傳遞資料的示例性方法。

【0014】 圖 5 圖示了用於從增強型記憶體模組上的計算核心記憶體元件傳遞資料的示例性方法。

【0015】 圖 6 圖示了用於利用增強型記憶體模組上的計算核心處理資料的示例性方法。

【0016】 圖 7 圖示了用於構造增強型記憶體模組的示例性方法。

【0017】 圖 8 圖示了用於包括增強型記憶體模組的處理系統的示例性記憶體映射。

【0018】 圖 9 圖示了用於利用包括計算核心的增強型記憶體模組來處理資料的示例性方法。

【0019】 圖 10 圖示了可以被配置為利用增強型記憶體模組執行資料處理的示例性電子設備。

【0020】 為了促進理解，在可能的情況下，已使用相同的元件符號來指定各圖所共有的元件。預期一個實施例的元件和特徵可以被有益地併入其他實施例而無需進一步的陳述。

【實施方式】

【0021】 本案的態樣提供了用於向諸如電腦、伺服器和其他電腦處理設備的處理系統中的記憶體硬體添加附加的機器學習和人工智慧計算能力的系統和方法。

【0022】 習知的電腦處理系統可以包括經由諸如 DIMM 介面的介面連接到諸如 DRAM 模組的記憶體模組並且經由諸

如 P C I e 介面的其他介面連接到其他周邊設備的處理器。隨著對機器學習和人工智慧「加速器」需求的增加，用於現有電腦處理系統的一種解決方案是將此類加速器連接到可用的系統介面，諸如可用的 P C I e 插槽。然而，此類周邊介面插槽的數量在任何給定的系統上皆可能受到限制，並且特別地，習知的電腦處理系統通常可以包括比 P C I e 插槽更多的 D I M M 插槽。

【0023】 本文描述的是包括整合加速器的增強型記憶體模組，該等整合加速器被配置為執行機器學習和人工智慧任務，諸如訓練和推理。記憶體整合加速器可以例如被實現於使用用於資料和指令通訊的標準記憶體介面的行業標準的雙列直插式記憶體模組（D I M M）中，諸如低負載 D I M M（L R D I M M）或帶暫存器的 D I M M（R D I M M）中。在某些實施例中，增強型記憶體模組可以被稱為例如「M L - D I M M」、「A I - D I M M」或「推理 D I M M」。

【0024】 在各種實施例中，記憶體整合加速器包括一或多個處理核心並且可以亦被配置為晶片上系統（S o C），其被整合到增強型記憶體模組中。處理核心能夠在習知儲存和取得之外對資料進行記憶體上處理。具體地，處理核心可以被配置為執行記憶體上的機器學習和人工智慧任務。

【0025】 此外，本文描述的增強型記憶體模組可以被配置為在主機處理系統（如在標準記憶體模組配置中）與用於記憶體上的機器學習和人工智慧任務的記憶體整合加速器之間分割和分配每個增強型記憶體模組中的記憶體元件及

/或記憶體空間。在某些實施例中，韌體元件可以在啟動時間（boot time）配置主機處理與記憶體上處理之間的記憶體的分配。

【0026】 對增強型記憶體模組中的記憶體的分割亦可以由安裝在包括增強型記憶體模組的設備上的程式、應用程式動態地配置，使得可以針對不同的任務和效能需求調整整體處理系統效能。有益的是，增強型記憶體模組可以被配置為充當正常記憶體模組，最大化對於主機處理系統的可用記憶體的容量，或者可以被配置為將記憶體中的某些或全部分配給記憶體整合加速器，最大化主機處理系統的機器學習和人工智慧計算容量。

【0027】 本文描述的實施例因此有益地使得附加的計算資源能夠被處理系統用於計算任務，諸如機器學習和人工智慧任務，而不會由於缺乏可用介面（例如，PCIe）而移除或阻止對其他必要元件的使用，此舉無需添加或修改附加的實體介面，亦無需實現全新的介面。附加的計算資源大大增加了現有處理系統以及新處理系統的容量。此外，本文描述的實施例利用雙操作模式架構，其有益地允許使用者在不同情況下選擇配置不同的操作模式以最大化處理系統的效用。

【0028】 本文描述的實施例可以被用於許多上下文中。例如，採用伺服器的資料中心可以實現本文描述的實施例以在具有相同數量的計算容量的同時採購更少的伺服器，此舉節省了設備成本、電力和冷卻成本、空間成本，並且亦

提供了更加功率高效的處理中心。作為另一實例，桌面及/或消費級計算系統可以實現本文描述的實施例以給家用電腦帶來改良的資料處理能力，而不是依賴於外部資料處理能力（例如，由雲端處理服務提供）。作為又一實例，物聯網路（IoT）設備可以實現本文描述的實施例以將資料處理能力擴展到全新類型的設備中，並進一步使得能夠在「邊緣」處進行處理（例如，機器學習和人工智慧推理）。

具有記憶體整合處理核心的示例性增強型記憶體模組

【0029】 圖1圖示了包括記憶體整合加速器104的增強型記憶體模組102的實例。

【0030】 在某些實施例中，增強型記憶體模組102根據標準化的形狀因數（`form factor`）而被構建，諸如LR-DIMM形狀因數，或者用於記憶體模組的另一標準化形狀因數。將標準化形狀因數用於增強型記憶體模組102允許在不修改底層系統介面或記憶體存取協定的情況下將其整合到現有處理系統中。在此實例中，記憶體模組102包括用於與處理系統中的DIMM插槽介面連接（`interface`）的引腳116。

【0031】 記憶體模組102包括複數個記憶體元件112A-F，其例如可以是動態隨機存取記憶體（DRAM）、同步DRAM（SDRAM）、低功率雙倍資料速率（LPDDR）DRAM、高頻寬記憶體（HBM）、圖形DDR（GDDR）記憶體等。在某些實例中，記憶體元件112A-F可以包括不同類型的記憶體元件的混合。

【0032】 記憶體元件 112 A - F 中的每一個被連接到鄰近記憶體緩衝器 114 A - F，鄰近記憶體緩衝器 114 A - F 被用於緩衝由主機處理系統（未圖示）從記憶體元件 112 A - F 讀取的資料以及向記憶體元件 112 A - F 寫入的資料。

【0033】 記憶體元件 112 A - F 通常可以被稱為雙模式或雙用途記憶體元件，因為其記憶體可以全部或部分地被分配給用於執行記憶體上處理的記憶體整合加速器 104 及 / 或用於執行習知系統記憶體功能的主機處理系統。圖 3 A - 圖 3 C 描述了不同記憶體分配的若干實例。

【0034】 記憶體整合加速器 104 包括被配置為儲存韌體的快閃記憶體 106，該韌體可以被配置為初始化記憶體模組 102。例如，韌體可以被配置為初始化（一或多個）計算核心 108 和增強型記憶體模組上的任何其他邏輯。另外，記憶體控制器（未圖示）可以被配置為初始化記憶體元件 112 A - F 以供（一或多個）計算核心 108 及 / 或主機處理系統使用。

【0035】 在某些實施例中，記憶體元件 112 A - F 的初始化可以由加速器 104 處理，此舉可以改良主機處理系統的啟動時間（例如，經由減少啟動時間）。在此類情況下，主機處理系統的記憶體驅動器僅需要初始化記憶體介面。

【0036】 在某些實施例中，快閃記憶體模組 106 中的韌體可以在啟動時間配置記憶體模組 102 並且將記憶體容量（例如，包括記憶體元件及 / 或記憶體範圍、位址等）的某些或全部分配（或映射）給加速器 104。快閃記憶體模組

106 中的韌體可以被更新，使得分配給加速器 104 的記憶體的量可以在運行時間（`runtime`）之間改變。在其中來自記憶體元件 112 A-F 的記憶體（例如，一定範圍的記憶體位址）在啟動時間被分配的實施例中，合併處理系統（`incorporating processing system`）將僅「看到」剩餘的可用記憶體（不包括分配給加速器 104 的記憶體）而不尋求向分配給加速器 104 的記憶體寫入，從而避免記憶體衝突。

【0037】 在某些實施例中，快閃記憶體模組 106 中的韌體可以在啟動時間配置初始記憶體分配，該初始記憶體分配隨後可以由來自合併處理系統的指令動態地改變。在此類情況下，合併處理系統的作業系統可以被配置為動態地重分配記憶體以供處理系統用於避免衝突。

【0038】 加速器 104 亦包括（一或多個）計算核心 108，其可以被配置為執行各種類型的處理任務，包括機器學習和人工智慧任務，如訓練和推理。在某些實施例中，計算核心 108 中的一或多個可以基於 ARM™ 架構、精簡指令集電腦（RISC）架構或複雜指令集架構電腦（CISC），在此僅舉若干實例。

【0039】 在某些實施例中，（一或多個）計算核心 108 替換並整合習知記憶體模組上的註冊表時鐘驅動器晶片或電路的功能。

【0040】 加速器 104 亦包括複數個記憶體元件 110 A-B，其專用於（一或多個）計算核心 108，或者換言之，單模

式或單用途的記憶體元件。由於記憶體元件 110 A - B 專用於（一或多個）計算核心 108，因此與記憶體元件 110 A - B 相關聯的記憶體不可由主機處理系統定址。專用記憶體元件 110 A - B 允許加速器 104 始終能夠執行計算任務。在某些實例中，記憶體元件 110 A - B 可以被配置為載入用於機器學習模型的模型參數，諸如權重、偏差等，該機器學習模型被配置為執行機器學習或人工智慧任務，諸如訓練或推理。

【0041】 在某些實施例中，加速器 104 可以被實現為與記憶體模組 102 整合的晶片上系統（SOC）。

【0042】 值得注意的是，圖 1 中每種類型的元件（例如，記憶體、緩衝器、快閃記憶體、計算核心等）的數量僅僅是一個實例，並且許多其他配置是可能的。此外，儘管在記憶體模組 102 的單側上圖示所有元件，但在其他實施例中，記憶體模組 102 可以在相對側上包括附加元件，諸如附加記憶體元件、緩衝器、計算核心、快閃記憶體模組等。

包括增強型記憶體模組的示例性處理系統

【0043】 圖 2 圖示了包括諸如參考圖 1 所描述的 102 的增強型記憶體模組的示例性處理系統 200。

【0044】 在此實例中，處理系統 200 實現包括作業系統（OS）使用者空間 202、OS 核心空間 212 和 CPU 218 的軟體堆疊。軟體堆疊被配置為與增強型記憶體模組 224 A 和 224 B 介面連接。

【0045】 通常，作業系統可以被配置為在諸如 202 的 OS 使用者空間與諸如 212 的 OS 核心空間之間分隔虛擬記憶體，以便提供針對惡意或錯誤的軟體行為的記憶體保護和硬體保護。在此實例中，OS 使用者空間 202 是其中應用軟體執行的記憶體區域，應用軟體諸如是 ML/AI 應用軟體 208 和加速器運行時間應用軟體 210。而核心空間 212 是為執行例如特權作業系統核心、核心擴展和裝置驅動程式（諸如記憶體驅動器 214 和加速器驅動器 216）而保留的記憶體空間。

【0046】 此外在此實例中，OS 使用者空間 202 包括機器學習及/或人工智慧應用軟體 208，其可以實現被配置為處理資料 206 的一或多個模型 204（例如，機器學習模型，諸如人工神經網路模型）。（一或多個）模型 204 可以包括例如權重、偏差和其他參數。資料 206 可以包括例如視訊資料、圖像資料、音訊資料、文字資料和其他類型的資料，（一或多個）模型 204 可以在該等資料上操作以執行各種機器學習任務，諸如圖像辨識和分割、語音辨識以及翻譯、價值預測等。

【0047】 ML/AI 應用軟體 208 亦被配置為與加速器運行時間應用軟體 210 介面連接，加速器運行時間應用軟體 210 是使用者空間過程，其被配置為將資料處理請求導引至處理加速器，諸如在增強型記憶體模組 224 A - B 中發現的彼等。值得注意的是，儘管圖 2 中圖示了單個 ML/AI 應用軟體 208，但此情形僅僅是一個實例，並且其他類型的處理

應用軟體可以存在於OS使用者空間202中並且被配置為利用增強型記憶體模組224A-B。

【0048】 加速器運行時間應用軟體210被配置為與OS核心空間212中的加速器驅動器216介面連接。通常，OS核心空間212中的驅動器是被配置為控制附接到處理系統（諸如處理系統200）的特定類型的設備的軟體。例如，驅動器可以向硬體設備（諸如記憶體管理單元220）提供軟體介面，此舉使得作業系統和其他電腦程式能夠存取硬體設備的功能，而無需知道關於硬體設備的精確細節。

【0049】 在此實例中，加速器驅動器216被配置為經由記憶體驅動器214提供對增強型記憶體模組224A-B的存取。在某些實施例中，加速器驅動器216被配置為執行由增強型記憶體模組224A-B使用的協定。為此，加速器驅動器216可以與記憶體驅動器214互動以向增強型記憶體模組224A-B發送命令。例如，此類命令可以被用於：將資料從分配給主機處理系統的記憶體移動到分配給增強型記憶體模組224A-B的加速器（例如，計算核心）的記憶體；開始處理資料；或者將資料從分配給加速器的記憶體移動到分配給主機處理系統的記憶體。

【0050】 在某些實施例中，加速器驅動器216可以實現使用特定記憶體位址來向增強型記憶體模組224A-B的記憶體寫入或從增強型記憶體模組224A-B的記憶體讀取的協定。在某些情況下，記憶體位址可以被實體映射到增強型記憶體模組。加速器驅動器216亦可以存取增強型記憶體

模組 2 2 4 A - B 上的 O S 使用者空間 2 0 2 不可存取的記憶體位址，其可以被用於將資料複製到分配給增強型記憶體模組 2 2 4 A - B 的加速器的記憶體中，並且隨後複製到彼等增強型記憶體模組的計算核心中。

【0051】 此外，加速器驅動器 2 1 6 被配置為與記憶體驅動器 2 1 4 介面連接。記憶體驅動器 2 1 4 被配置為提供對增強型記憶體模組 2 2 4 A - B 的記憶體元件的低級存取。例如，記憶體驅動器 2 1 4 可以被配置為將加速器驅動器 2 1 6 產生的命令直接傳遞到增強型記憶體模組 2 2 4 A - B。記憶體驅動器 2 1 4 可以是作業系統記憶體管理的部分，但可以被修改以啟用本文描述的增強型記憶體模組功能。例如，記憶體驅動器 2 1 4 可以被配置為在其中資料正在增強型記憶體模組 2 2 4 A - B 內的分配給加速器（例如，計算核心）的記憶體與分配給主機處理系統的記憶體之間被複製的時段期間，阻止對增強型記憶體模組 2 2 4 A - B 的存取。

【0052】 記憶體驅動器 2 1 4 被配置為與記憶體管理單元（M M U）2 2 0 介面連接，M M U 2 2 0 是被配置為管理與諸如 C P U 2 1 8 的處理器相關聯的記憶體和快取操作的硬體元件。

【0053】 記憶體管理單元 2 2 0 被配置為與記憶體控制器 2 2 2 介面連接，記憶體控制器 2 2 2 是被配置為管理進出主機處理系統的主記憶體的資料流程的硬體元件。在某些實施例中，記憶體控制器 2 2 2 是單獨的晶片或者被整合到另一晶片中，諸如是微處理器（如 C P U 2 1 8）的組成部分，（在

此種情況下，其可以被稱為整合記憶體控制器（IMC）。記憶體控制器222可以替代地被稱為記憶體晶片控制器（MCC）或記憶體控制器單元（MCU）。

【0054】 在此實例中，記憶體控制器222被配置為控制進出增強型記憶體模組224A-B的資料流程。在某些實施例中，增強型記憶體模組224A-B可以是增強型DIMM，如前述。

【0055】 增強型記憶體模組224A-B分別包括驅動器226A-B，其使得能夠與處理系統200的元件進行通訊。例如，驅動器226A-B可以經由與記憶體控制器222的DIMM介面而使得能夠與處理系統200的其他元件進行通訊。

【0056】 通常而言，ML/AI應用軟體208可以經由加速器運行時間應用軟體210請求經由將由增強型記憶體模組執行的模型204對資料206進行處理，加速器運行時間應用軟體210將請求傳達到加速器驅動器216並且隨後傳達到記憶體驅動器214上，記憶體驅動器214可以是另外的習知記憶體驅動器。記憶體驅動器214隨後可以經由記憶體管理單元220和記憶體控制器222向增強型記憶體模組224A-B中的一或多個路由處理指令和資料。此情形僅是處理流程的一個實例，其他實例亦是可能的。

【0057】 在某些實施例中，若多個增強型記憶體模組（例如，224A-B）正被用於相同的應用軟體（例如，ML/AI應用軟體208），則加速器運行時間應用軟體210和加速器

驅動器 216 可以協調以管理跨系統的資料。例如，若單個增強型記憶體模組不能完成全部的工作負載，加速器運行時間應用軟體 210 可以從加速器驅動器 216 請求總共可用的加速器分配的記憶體（以及增強型記憶體模組的數量），並且隨後則儘可能高效地在增強型記憶體模組之間分割工作負載。作為另一實例，在多個工作負載和多個增強型記憶體模組的情況下，加速器運行時間應用軟體 210 可以被配置為將工作負載分割到可用的增強型記憶體模組以最大化效能。

【0058】 CPU 218 通常可以是包括其自己的處理核心中的一或多個的中央處理單元。儘管在此實例中 CPU 218 不執行被路由到增強型記憶體模組 224 A - B 的來自 ML / AI 應用軟體 208 的特定處理請求，但在其他實例中 CPU 218 可以與增強型記憶體模組 224 A - B 協同地對來自 ML / AI 應用軟體 208（和其他應用軟體）的請求執行並行處理。

【0059】 此外，儘管已經針對此實例圖示和描述了處理系統 200 的某些態樣，但是處理系統 200 可以包括其他元件。例如，處理系統 200 可以包括附加的應用程式、附加的硬體元件（例如，如參考圖 10 所描述的其他種類的處理單元）等等。圖 2 有意地聚焦於與所描述的實施例相關的處理系統 200 的某些元件上。

增強型記憶體模組的示例性操作模式

【0060】 圖 3 A - 圖 3 C 圖示了用於增強型記憶體模組（例如，圖 1 中的 1 0 2 和圖 2 中的 2 2 4 A - B）的示例性記憶體分配配置。

【0061】 在圖 3 A 中，增強型記憶體模組 3 0 2 例如在啟動時間由快閃記憶體 3 0 6 中的韌體配置為僅將其專用記憶體元件 3 1 0 A - B 用於由（一或多個）計算核心 3 0 8 進行的任何處理。在此配置中，增強型記憶體模組 3 0 2 被設置為向系統記憶體提供最大容量。

【0062】 在圖 3 B 中，增強型記憶體模組 3 0 2 被配置為將其專用記憶體元件 3 1 0 A - B 加上由記憶體元件 3 1 2 A - C 提供的記憶體空間用於由（一或多個）計算核心 3 0 8 進行的任何處理。另一態樣，與習知系統記憶體一樣，由記憶體元件 3 1 2 D - F 提供的記憶體空間是為主機處理系統保留的。因此，在此配置中，增強型記憶體模組 3 0 2 被設置為平衡模組上計算能力與主機處理系統記憶體需求。

【0063】 在圖 3 C 中，增強型記憶體模組 3 0 2 被配置為將其專用記憶體元件 3 1 0 A - B 加上由記憶體元件 3 1 2 A - F 提供的記憶體空間用於由（一或多個）計算核心 3 0 8 進行的任何處理。因此，在此配置中，增強型記憶體模組 3 0 2 被設置為最大化模組上計算能力。值得注意的是，在此類配置中，另一增強型記憶體模組或另一習知記憶體模組可以提供用於主機處理系統操作的主機處理系統可定址記憶體。

【0064】 值得注意的是，儘管在該實例中各種記憶體元件被圖示為分配給（一或多個）計算核心 3 0 8 或主機系統記

記憶體，但該表示是為了方便。例如，虛擬記憶體位址空間可以被分配給記憶體上計算或主機系統處理，而不管彼等位址空間如何被分配給實體記憶體元件內的實體記憶體位址（例如，經由頁表）。

【0065】 此外，圖 3 A - 圖 3 C 中圖示的示例性配置僅是一些示範性的可能情況。由增強型記憶體模組 3 0 2 提供並分配給（一或多個）計算核心 3 0 8 的主機處理系統記憶體空間的量可以佔可用的、主機處理系統可定址記憶體元件的 0 與 1 0 0 % 之間的任何比率。然而，在某些實施例中，分配給（一或多個）計算核心 3 0 8 的記憶體空間的比率可能受到限制，使得每個增強型記憶體模組上的至少某些記憶體空間被保留用於主機系統。

用於向增強型記憶體模組上的計算核心記憶體元件傳遞資料的示例性方法

【0066】 圖 4 圖示了用於向增強型記憶體模組（例如，圖 1 中的 1 0 2、圖 2 中的 2 2 4 A - B 和圖 3 A - C 中的 3 0 2）上的計算核心記憶體元件傳遞資料的示例性方法。

【0067】 方法 4 0 0 開始於步驟 4 0 2，其中經由增強型記憶體協定接收資料命令以將資料移動到增強型記憶體模組上的計算核心記憶體元件中。在某些實施例中，諸如上文參考圖 2 所論述的加速器驅動器可以經由增強型記憶體協定產生命令。

【0068】 注意，在此實例中，計算核心記憶體元件可以是指定的計算核心記憶體元件（諸如圖 3 中的記憶體元件

3 1 0 A - B) 或者已經被分配給計算核心的雙模式記憶體元件 (諸如圖 3 中的記憶體元件 3 1 2 A - C) 。此外，方法 4 0 0 是參考將資料移入和移出記憶體元件來描述的，但此舉是為了便於演示。分配給計算核心的記憶體可以是記憶體位址的範圍、記憶體空間或任何其他邏輯，以及實體記憶體子部分。

【0069】 方法 4 0 0 隨後進行到步驟 4 0 4 ，其中阻止用於存取增強型記憶體模組上的記憶體元件的主機處理系統命令。此舉是為了防止在利用增強型記憶體模組上的計算核心處理資料命令時發生任何衝突。

【0070】 在某些實施例中，諸如上文參考圖 2 所描述的記憶體驅動器可以被配置為阻止主機處理系統命令。

【0071】 方法 4 0 0 隨後進行到步驟 4 0 6 ，其中在資料匯流排上設置忙狀態以在資料在增強型記憶體模組上被移動時阻止來自主機處理系統的讀取請求。此舉是為了防止在利用增強型記憶體模組上的計算核心處理資料命令時發生任何中斷。

【0072】 方法 4 0 0 隨後進行到步驟 4 0 8 ，其中從分配給主機處理系統的記憶體元件向分配給增強型記憶體模組上的計算核心的記憶體元件傳遞資料。如前述，分配給計算核心的記憶體元件可以包括指定的計算核心記憶體元件 (諸如圖 3 中的記憶體元件 3 1 0 A - B) 或者已經被分配給計算核心的雙模式記憶體元件 (諸如圖 3 中的記憶體元件 3 1 2 A - C) 。此處，傳遞到計算核心記憶體元件中的資料

可以包括用於處理的資料以及用於如何處理資料的指令。例如，資料可以包括用於處理的資料（例如，圖2中的206）、模型資料（例如，圖2中的204）和處理命令。

【0073】 方法400隨後進行到步驟410，其中決定資料傳遞是否完成。若不是，則該方法返回到步驟408並繼續資料傳遞。

【0074】 若傳遞在步驟410處完成，則方法400進行到步驟412，其中忙狀態被從資料匯流排清除。

【0075】 方法400繼續進行到步驟414，其中解除阻止（unblock）對分配給主機處理系統的記憶體元件的主機處理系統命令，使得利用主機處理系統的正常記憶體操作可以繼續。

【0076】 方法400隨後進行到步驟416，其中利用增強型記憶體模組的一或多個計算核心來處理資料，諸如下文參考圖6更詳細地描述的一般。例如，處理資料可以包括執行機器學習任務，諸如訓練或推理。

用於從增強型記憶體模組上的計算核心記憶體元件傳遞資料的示例性方法

【0077】 圖5圖示了用於從增強型記憶體模組（例如，圖1中的102、圖2中的224A-B和圖3A-C中的302）上的計算核心傳遞資料的示例性方法500。例如，方法500可以在圖4的步驟416處完成對資料的處理之後被執行。

【0078】 方法500開始於步驟502，其中經由增強型記憶體協定接收結果命令以將資料移出增強型記憶體模組上的

計算核心記憶體元件。注意，在此實例中，計算核心記憶體元件可以是指定的計算核心記憶體元件（諸如圖3中的記憶體元件310A-B）或者已經被分配給計算核心的雙模式記憶體元件（諸如圖3中的記憶體元件312A-C）。

【0079】 方法500隨後進行到步驟504，其中阻止用於存取增強型記憶體模組上的記憶體元件的主機處理系統命令。此舉是為了防止在增強型記憶體模組上處理資料命令時發生任何衝突。

【0080】 方法500隨後進行到步驟506，其中在資料匯流排上設置忙狀態以在資料在增強型記憶體模組上被移動時阻止來自主機處理系統的讀取請求。

【0081】 方法500隨後進行到步驟508，其中從計算核心記憶體元件向增強型記憶體模組上被分配給主機處理系統的記憶體元件傳遞資料。如前述，分配給計算核心的記憶體元件可以包括指定的計算核心記憶體元件（諸如圖3中的記憶體元件310A-B）或者已經被分配給計算核心的雙模式記憶體元件（諸如圖3中的記憶體元件312A-C）。

【0082】 方法500隨後進行到步驟510，其中決定傳遞是否完成。若不是，則該方法返回到步驟508並繼續資料傳遞。

【0083】 若傳遞在步驟510處完成，則方法500進行到步驟512，其中忙狀態被從資料匯流排清除。

【0084】 最終，方法500繼續進行到步驟514，其中解除阻止對分配給主機處理系統的記憶體元件的主機處理系統命令，使得利用主機處理系統的正常記憶體操作可以繼續。
用於利用增強型記憶體模組上的計算核心處理資料的示例性方法

【0085】 圖6圖示了用於利用增強型記憶體模組上的計算核心處理資料的示例性方法600。例如，利用增強型記憶體模組（例如，圖1中的102、圖2中的224A-B和圖3A-C中的302）的計算核心處理資料可以在向增強型記憶體模組上的計算核心記憶體元件傳遞資料之後發生，諸如上文參考圖4所描述的一般。用於處理的資料可以由應用程式、感測器、其他處理器等產生。例如，用於處理的資料可以包括圖像、視訊或聲音資料。

【0086】 方法600開始於步驟602，其中從請求者接收用於在增強型記憶體模組上處理資料的命令。例如，請求者可以是資料處理應用軟體，諸如圖2中的ML/AI應用軟體208。

【0087】 在某些實施例中，請求者查詢加速器運行時間應用軟體（例如，圖2中的210）以決定有多少增強型記憶體模組在系統中被配置，以及彼等增強型記憶體模組上有多少記憶體被分配為執行基於加速的工作負載。加速器運行時間應用軟體隨後可以查詢加速器驅動器（例如，圖2中的216）以得到此資訊。利用此資訊，請求者可以決定工作負載是否適合所配置的記憶體量並且相應地執行該工作負

載。在可用記憶體的量不足以用於工作負載的情況下，請求者可以重配置工作負載，例如減小其大小或將工作負載分塊，使得可用記憶體可以被利用。

【0088】 方法600隨後進行到步驟604，其中將資料載入到增強型記憶體模組中的（一或多個）計算核心記憶體元件中。如前述，（一或多個）計算核心記憶體元件可以是專用記憶體元件或動態分配的記憶體元件。在某些實施例中，步驟604根據參考圖4描述的方法400而被執行。

【0089】 在某些實施例中，請求者（例如，圖2中的ML/AI應用軟體208）將向加速器運行時間應用軟體（例如，圖2中的210）發送資料，該加速器運行時間應用軟體隨後向加速器驅動器（例如，圖2中的216）轉發資料。加速器驅動器隨後可以向記憶體驅動器（例如，圖2中的214）發送用於載入資料的協定命令，並且記憶體驅動器隨後將向增強型記憶體模組（圖2中的224A-B）發送該命令。

【0090】 方法600隨後進行到步驟606，其中經由增強型記憶體協定接收處理命令，諸如下文更詳細地描述的一般。處理命令可以被配置為使得增強型記憶體模組的（一或多個）處理核心對儲存在計算核心記憶體元件中的資料執行資料處理。

【0091】 在某些實施例中，在載入資料之後，加速器驅動器（例如，圖2中的216）將向記憶體驅動器（例如，圖2中的214）發送用於處理資料的協定命令，該記憶體驅動

器隨後將向增強型記憶體模組（例如，圖2中的224A-B）轉發該命令以供執行。

【0092】 方法600隨後進行到步驟608，其中由增強型記憶體模組上的（一或多個）計算核心處理儲存在計算核心記憶體元件中的資料。在某些實施例中，處理可以關於機器學習及/或人工智慧任務，諸如訓練或推理。此外，在某些實施例中，機器學習任務可以跨單個增強型記憶體模組上的複數個計算核心以及跨複數個增強型記憶體模組而被並行處理。

【0093】 方法600隨後進行到步驟610，其中若工作負載的處理尚未完成，則方法600返回到步驟608。

【0094】 若在步驟610處工作負載的處理完成，則方法600進行到步驟612，其中結果信號經由增強型記憶體協定而被發送。

【0095】 在某些實施例中，加速器驅動器（例如，圖2中的216）可以定期地查詢增強型記憶體模組（例如，圖2中的224A-B）以決定處理何時完成。當處理完成時，增強型記憶體模組可以在匯流排上輸出指示處理完成的信號。隨後，在下次加速器驅動器查詢增強型記憶體模組時，該加速器驅動器將觀察到結果信號。

【0096】 方法600隨後進行到步驟612，其中經處理的資料從計算核心記憶體元件被移動至增強型記憶體模組上被分配給主機處理系統的記憶體元件。例如，參考圖3B，資料可以從分配給計算核心的記憶體元件310A-B和

3 1 2 A - C 中 的 一 或 多 個 被 移 動 到 分 配 給 主 機 處 理 系 統 的 記 憶 體 元 件 3 1 2 D - F 中 的 一 或 多 個 。

【0097】 在某些實施例中，在處理完成之後，加速器驅動器（例如，圖 2 中的 2 1 6）將向記憶體驅動器（例如，圖 2 中的 2 1 4）發送協定命令以取得資料。記憶體驅動器隨後將向增強型記憶體模組傳遞該命令，增強型記憶體模組進而將經處理的資料複製到主機處理系統可定址記憶體中。

【0098】 方法 6 0 0 隨後在步驟 6 1 4 處結束，其中經處理的資料從增強型記憶體模組上被分配給主機處理系統的記憶體元件而被提供給請求者。

【0099】 在某些實施例中，加速器驅動器（例如，圖 2 中的 2 1 6）將向記憶體驅動器（例如，圖 2 中的 2 1 4）發送命令以讀取出經處理的資料。記憶體驅動器隨後將向增強型記憶體模組轉發該等讀取請求。加速器驅動器將經由加速器運行時間應用軟體（例如，圖 2 中的 2 1 0）將該等結果發送回請求者。

增強型記憶體協定的示例性態樣

【0100】 增強型記憶體協定可以被實現以用於與增強型記憶體模組（例如，圖 1 中的 1 0 2 或圖 2 中的 2 2 4 A - B 或圖 3 A - C 中的 3 0 2）互動。在某些實施例中，增強型記憶體協定可以基於聯合電子設備工程委員會（Joint Electron Device Engineering Council，JEDEC）協定或者是其擴展。

【0101】 在某些實施例中，增強型記憶體協定可以實現用於與記憶體整合加速器（諸如參考圖1描述的加速器104）介面連接的各種命令。

【0102】 例如，增強型記憶體協定可以包括「初始化」信號（或命令），該「初始化」信號（或命令）被配置為從例如快閃記憶體模組（諸如上文參考圖1描述的模組106）載入韌體，並且初始化記憶體整合加速器。

【0103】 增強型記憶體協定亦可以包括「資料」信號（或命令），該「資料」信號（或命令）被配置為諸如經由參考圖4描述的方法而啟動用於將資料從增強型記憶體模組（例如，圖1中的102、圖2中的224A-B和圖3A-C中的302）上被分配給系統記憶體的記憶體元件（例如，圖3B中的312D-F）移動到計算核心記憶體元件（例如，圖3B中的312A-C）的過程，如前述，計算核心記憶體元件可以是靜態的或者是動態分配的。

【0104】 「資料」信號可以具有變體，諸如：「移動所有資料」，其中起始位址與結束位址之間的所有資料在系統分配的記憶體與計算核心分配的記憶體之間被移動；「移動 x 個位元組」，其中 x 個資料位元組從系統分配的記憶體被移動到計算核心分配的記憶體；及「移動 x 個週期」，其中 x 個週期將被用於將資料從系統分配的記憶體移動到計算核心分配的記憶體。

【0105】 增強型記憶體協定亦可以包括「處理」信號（或命令），該「處理」信號（或命令）被配置為諸如經由參

考圖6描述的方法而使得增強型記憶體模組（例如，圖1中的102、圖2中的224A-B和圖3A-C中的302）的計算核心開始執行已被儲存在計算核心分配的記憶體中的指令。

【0106】 增強型記憶體協定亦可以包括「結果」信號（或命令），該「結果」信號（或命令）被配置為諸如經由參考圖5描述的方法而啟動用於將資料從計算核心分配的記憶體（例如，圖3B中的312A-C）移動到系統分配的記憶體（例如，圖3B中的312D-F）的過程，使得其可以被主機處理系統（例如，圖2中的系統200）的其他部分存取。

【0107】 因此，在一個實例中，加速器驅動器（例如，圖2中的216）發送「資料」信號並且記憶體驅動器（例如，圖2中的214）向增強型記憶體模組（例如，圖2中的224A-B）轉發該信號。加速器驅動器告訴記憶體驅動器阻止訊務（traffic），此舉可以是帶有「資料」命令的單個命令或者單獨的命令。記憶體驅動器隨後會阻止並非來自加速器驅動器的所有訊務。在某些情況下，訊務可以被保持或者被發回以用於重試。增強型記憶體模組亦可以在緩衝器中設置值以輸出指示記憶體傳遞狀態的資料樣式。例如，加速器驅動器可以從增強型記憶體模組請求資料（記憶體驅動器轉發該請求），若資料樣式為「忙」，則加速器驅動器等待並再次嘗試，若資料樣式為「完成」，隨後加速器驅動器告訴記憶體驅動器再次允許訊務。

用於構造增強型記憶體模組的示例性方法

【0108】 圖 7 圖示了用於構造增強型記憶體模組（例如，圖 1 中的 1 0 2、圖 2 中的 2 2 4 A - B 和圖 3 A - C 中的 3 0 2）的示例性方法 7 0 0。

【0109】 方法 7 0 0 開始於步驟 7 0 2，其中將雙模式記憶體元件（諸如圖 1 中的增強型記憶體模組 1 0 2 上的記憶體元件 1 1 2 A - F）放置在增強型記憶體模組上。在此實例中，「雙模式」指的是每個記憶體元件的以習知方式被用於系統記憶體以及被用於針對（一或多個）模組上計算核心（諸如圖 1 中的 1 0 8）的本端記憶體的能力。該等記憶體元件無需實體修改即可在任一模式下工作。

【0110】 方法 7 0 0 隨後進行到步驟 7 0 4，其中將計算核心（諸如圖 1 中的（一或多個）計算核心 1 0 8）放置在增強型記憶體模組上。

【0111】 方法 7 0 0 隨後進行到步驟 7 0 6，其中在（一或多個）計算核心與每個雙模式記憶體元件之間連接命令/位址線。命令/位址線可以被配置為傳送用於決定記憶體的何者部分被存取以及決定命令是什麼的信號。

【0112】 方法 7 0 0 隨後進行到步驟 7 0 8，其中在每個雙模式記憶體元件與（一或多個）計算核心（用於本端處理模式）和記憶體緩衝器（用於系統記憶體模式）之間連接資料線。

【0113】 方法 7 0 0 隨後進行到步驟 7 1 0，其中將計算核心記憶體元件（諸如圖 1 中的增強型記憶體模組 1 0 2 上的記憶體元件 1 1 0 A - B）放置在增強型記憶體模組上。

【0114】 方法700隨後進行到步驟712，其中在（一或多個）計算核心與計算核心記憶體元件之間連接命令/位址線和資料線。

【0115】 方法700隨後進行到步驟714，其中將快閃記憶體模組（諸如圖1中的增強型記憶體模組102上的快閃記憶體模組106）放置在增強型記憶體模組上。

【0116】 方法700隨後進行到步驟716，其中將快閃記憶體模組連接到（一或多個）計算核心，諸如圖1中的增強型記憶體模組102的（一或多個）計算核心108。

【0117】 方法700隨後進行到步驟718，其中向快閃記憶體模組寫入韌體以初始化增強型記憶體模組，諸如圖1的增強型記憶體模組102。

【0118】 值得注意的是，方法700僅是如何構造增強型記憶體模組（諸如圖1的增強型記憶體模組102）的一個高級實例，亦存在很多其他的實例。為清楚起見，省略了對具體製造技術的論述。

用於包括增強型記憶體模組的處理系統的示例性記憶體映射

【0119】 圖8圖示了用於包括增強型記憶體模組（例如，圖1中的102、圖2中的224A-B和圖3A-C中的302）的處理系統的示例性記憶體映射。

【0120】 在所圖示的實例中，處理系統具有六個記憶體模組插槽（例如，6個DIMM插槽）並且每個DIMM插槽之每一者記憶體模組具有8GB的系統可定址記憶體，總共48

GB 的系統記憶體。每個 DIMM 插槽中的記憶體模組可以是習知記憶體模組或增強型記憶體模組，使得系統作為整體可以是「混合配置」——具有習知和增強型記憶體模組的混合，或者可以是「增強型配置」——所有皆是增強型記憶體模組。因為增強型記憶體模組可以被配置為在主機處理系統與記憶體上處理元件（例如，計算核心，諸如圖 1 中的 108）之間分配記憶體，所以系統中存在的增強型記憶體模組越多，可分配給記憶體上處理的總記憶體越多。

【0121】 此外，在此實例中，實現記憶體分配閾值，使得任何增強型記憶體模組中多達一半的記憶體可以被分配給記憶體上處理（例如，給計算核心，諸如圖 1 中的 108），並且剩餘的記憶體被分配給主機處理系統。在其他實施例中，記憶體分配閾值可以不被實現，或者可以用不同的值來實現。記憶體閾值通常可以被建立以確保主機處理系統維持足夠的所分配的記憶體以正確地執行。

【0122】 記憶體映射 802 或 804 圖示了具有記憶體模組的混合配置的處理系統中的示例性記憶體分配。記憶體映射 806 和 808 圖示了具有增強型配置的處理系統中的示例性記憶體分配，其中所有記憶體模組皆是增強型記憶體模組。在所圖示的實例中，806 和 808 的增強型配置中的記憶體分配包括分配給計算核心以用於記憶體上處理的更多記憶體。

【0123】 此外，記憶體映射 806 和 808 中的記憶體分配特定於每個增強型記憶體模組（例如，增強型記憶體模組 0-5）

的計算核心。然而，在其他實例中，記憶體通常可以作為集合被分配給計算核心，諸如在示例性記憶體映射802和804中，而不是具體地，如在示例性記憶體映射806和808中。

【0124】 通常而言，具有更多增強型記憶體模組增加了可以被分配給記憶體上處理的記憶體的量，從而改良了主機處理系統執行例如AI和ML特定任務（諸如訓練和推理）的能力，以及增加了配置總系統記憶體的靈活性。

【0125】 在主機處理系統與記憶體上處理（諸如，經由像圖1的104一樣的加速器）之間對記憶體進行動態重分配的情況下，記憶體映射可以被相應地更新。例如，記憶體映射802可以被更新為記憶體映射806或其他組合。

【0126】 圖8中圖示的實例意欲說明使用增強型記憶體模組的各種記憶體配置，但僅僅是說明若干實例，其他實例亦是可能的。

用於利用增強型記憶體模組上處理資料的示例性方法

【0127】 圖9圖示了用於利用包括計算核心（例如，圖1中的108）的增強型記憶體模組（例如，圖1中的102、圖2中的224A-B和圖3A-C中的302）來處理資料的示例性方法900。在某些實施例中，計算核心可以是記憶體整合加速器的部分，諸如圖1的104。在某些實施例中，計算核心可以包括複數個處理元件，諸如複數個單獨的處理核心。在某些實施例中，增強型記憶體模組包括低負載雙列直插式記憶體模組（LRDIMM）。

【0128】 方法 900 開始於步驟 902，其中經由將增強型記憶體模組的記憶體的第一子集分配為主機處理系統可定址記憶體並將增強型記憶體模組的記憶體的第二子集分配為計算核心可定址記憶體來初始化增強型記憶體模組，諸如例如參考圖 3A - 圖 3C 所描述的一般。

【0129】 在方法 900 的某些實施例中，增強型記憶體模組的初始化可以在啟動時間被執行。在其他實施例中，增強型記憶體模組的初始化可以在主機處理系統的操作期間被動態地執行。

【0130】 在某些實施例中，記憶體的分配可以包括在計算核心與主機處理系統之間分配一或多個雙模式記憶體元件（諸如圖 1 中的記憶體元件 112A - F）的記憶體。例如，分配可以包括一或多個記憶體範圍或記憶體位址空間，諸如虛擬位址空間或實體位址空間。

【0131】 方法 900 繼續進行到步驟 904，其中在增強型記憶體模組處從主機處理系統接收資料。資料可以由主機處理系統提供給增強型記憶體模組以用於在增強型記憶體模組上進行處理，諸如由計算核心處理。

【0132】 在某些實施例中，所接收的資料可以包括用於由機器學習模型處理的資料和機器學習模型參數。在某些實施例中，所接收的資料亦可以包括處理命令或指令。

【0133】 方法 900 隨後進行到步驟 906，其中將所接收的資料儲存在記憶體的第一子集（主機處理系統可定址記憶體）中。在某些實施例中，主機處理系統可定址記憶體包

括增強型記憶體模組上的一或多個雙模式記憶體元件（諸如圖1中的記憶體元件112A-F）中分配給主機處理系統的記憶體空間。

【0134】 方法900隨後進行到步驟908，其中從記憶體的第一子集（主機處理系統可定址記憶體）向記憶體的第二子集（計算核心可定址記憶體）傳遞所接收的資料。

【0135】 在某些實施例中，計算核心可定址記憶體包括非主機處理系統可定址記憶體，諸如圖1中的記憶體元件110A-B。在某些實施例中，計算核心可定址記憶體僅包括非主機處理系統可定址記憶體（諸如在圖3A的實例中）。

【0136】 方法900隨後進行到步驟910，其中利用增強型記憶體模組上的計算核心處理所接收的資料以產生經處理的資料。

【0137】 方法900隨後進行到步驟912，其中從記憶體的第二子集（計算核心可定址記憶體）向記憶體的第一子集（主機處理系統可定址記憶體）傳遞經處理的資料。

【0138】 方法900隨後進行到步驟914，其中經由記憶體的第一子集（主機處理系統可定址記憶體）向主機處理系統提供經處理的資料，諸如上文參考圖2所描述的一般。

【0139】 在某些實施例中，初始化增強型記憶體模組包括處理儲存在增強型記憶體模組上的快閃記憶體模組中的韌體指令，諸如上文參考圖1所描述的一般。

【0140】 在某些實施例中，方法900亦包括以下步驟：在從記憶體的第一子集（主機處理系統可定址記憶體）向記

憶體的第二子集（計算核心可定址記憶體）傳遞資料之前，在主機處理系統的資料匯流排上設置忙狀態以指示資料的傳遞正在發生並阻止對主機處理系統可定址記憶體的讀取請求，諸如上文參考圖4所描述的一般。

【0141】 在某些實施例中，方法900亦包括以下步驟：在從記憶體的第一子集（主機處理系統可定址記憶體）向記憶體的第三子集（計算核心可定址記憶體）傳遞資料之後，在主機處理系統的資料匯流排上設置可用狀態以允許從主機處理系統向主機處理系統可定址記憶體的讀取請求或寫入請求，諸如上文參考圖4所描述的一般。

【0142】 在某些實施例中，方法900亦包括以下步驟：在從記憶體的第一子集（主機處理系統可定址記憶體）向記憶體的第三子集（計算核心可定址記憶體）傳遞資料之前，啟用對用於增強型記憶體模組的主機處理系統記憶體命令的阻止，諸如上文參考圖4所描述的一般。

【0143】 在某些實施例中，方法900亦包括以下步驟：在從記憶體的第一子集（主機處理系統可定址記憶體）向記憶體的第三子集（計算核心可定址記憶體）傳遞資料之後，禁用對用於增強型記憶體模組的主機處理系統記憶體命令的阻止，諸如上文參考圖4所描述的一般。

【0144】 在某些實施例中，方法900亦包括以下步驟：經由解除分配記憶體的第一子集和記憶體的第三子集，以及將一或多個雙模式記憶體元件中的記憶體的第四子集分配為主機處理系統可定址記憶體以及將一或多個雙模式記憶

體元件中的記憶體的第四子集分配為計算核心可定址記憶體來重分配增強型記憶體模組的記憶體，其中記憶體的第三子集不同於第一子集，並且其中記憶體的第四子集不同於第二子集。在某些實施例中，重分配在啟動時間之後（諸如當主機處理系統正在執行時）被執行。如此，重分配可以被稱為記憶體的動態或「熱」重分配。

【0145】 例如，在某些實施例中，第一分配（諸如圖 3 A 的分配）可以被動態地重分配為第二分配（像圖 3 B 或圖 3 C 一樣的分配，或其他未被圖示的分配）。在某些實施例中，主機處理系統在重分配過程期間可以被阻止向增強型記憶體模組發送記憶體命令。在某些實施例中，頁表或記憶體映射可以在重分配之後被更新，使得實體記憶體對主機處理系統不再可存取或最近可存取，主機處理系統被適當地配置。例如，圖 8 中的記憶體映射實例可以在增強型記憶體模組上的記憶體的重分配之後被更新。

用於利用增強型記憶體模組進行資料處理的電子設備實例

【0146】 圖 10 圖示了示例性電子設備 1000，其可以被配置為利用增強型記憶體模組來執行資料處理，諸如本文參考圖 4 - 圖 6 和圖 9 所描述的一般。在某些實施例中，電子設備 1000 可以包括伺服器電腦。

【0147】 電子設備 1000 包括中央處理單元（CPU）1002，其在某些實例中可以是多核 CPU。在 CPU 1002 處執行的指令可以例如從與 CPU 1002 相關聯的程式記憶體被載入或者可以從記憶體分區 1024 被載入。

【0148】 電子設備 1000 亦包括針對特定功能定製的附加處理元件，諸如圖形處理單元（GPU）1004、數位信號處理器（DSP）1006和神經處理單元（NPU）1008。

【0149】 NPU（諸如1008）通常是專用電路，其被配置用於實現針對執行機器學習演算法（諸如用於處理人工神經網路（ANN）、深度神經網路（DNN）、隨機森林（RF）的演算法等）的所有必要的控制和算術邏輯。NPU有時亦可以被稱為神經信號處理器（NSP）、張量處理單元（TPU）、神經網路處理器（NNP）、智慧處理單元（IPU）、視覺處理單元（VPU）或圖形處理單元。

【0150】 NPU可以針對訓練或推理而被最佳化，或者在某些情況下被配置為平衡兩者之間的效能。對於能夠執行訓練和推理兩者的NPU，該兩個任務通常仍舊可以被獨立地執行。

【0151】 例如，NPU（諸如1008）可以被配置為加速常見的機器學習推理任務的效能，該等任務諸如是圖像分類、機器翻譯、物件偵測以及各種其他預測模型。在某些實例中，複數個NPU可以在單個晶片上被產生實體，諸如晶片上系統（SoC），而在其他實例中，該等NPU可以是專用神經網路加速器的部分。被設計用於加速推理的NPU通常被配置為在完整的模型上操作。此類NPU因此可以被配置為輸入新的資料片段，並經由已訓練的模型對其進行快速處理以產生模型輸出（例如，推理）。

【0152】 作為另一實例，NPU可以被配置為加速常見的機器學習訓練任務，諸如處理測試資料集（常被標示或標記），在資料集上進行反覆運算（*iterate*），以及隨後調整模型參數，諸如權重和偏差，以便改良模型效能。通常，基於錯誤預測的最佳化涉及經由模型的層向回傳播，以及決定梯度以減少預測誤差。

【0153】 在一個實現中，NPU 1008是CPU 1002、GPU 1004及/或DSP 1006中的一或多個的部分。

【0154】 電子設備1000亦可以包括一或多個輸入及/或輸出設備1022，諸如螢幕、網路介面、實體按鈕等。

【0155】 電子設備1000亦包括記憶體1024，其代表一或多個靜態及/或動態的記憶體，諸如動態隨機存取記憶體、基於快閃記憶體的靜態記憶體等。在此實例中，記憶體1024包括可以由電子設備1000的前述處理器中的一或多個執行的電腦可執行元件。

【0156】 記憶體1024可以代表一或多個記憶體模組，諸如本文（例如，參考圖1、圖2和圖3）描述的增強型記憶體模組。例如，記憶體1024可以包括一或多個計算核心1024I和快閃記憶體元件1024J。在某些情況下，（一或多個）計算核心1024I可以包括NPU或如本文所述的其他種類的處理器。此外，儘管未圖示，但是記憶體1024中的某些可以僅由（一或多個）計算核心1024J存取，而某些記憶體可以被配置為可以由處理系統1000或計算核心1024J存取。

【0157】 在此實例中，記憶體1024包括接收元件1024A、儲存元件1024B、傳遞元件1024C、處理元件1024D、發送（或提供）元件1024E、初始化元件1024F、分配元件1024G和分割元件1024H。所圖示的元件和其他未圖示的元件可以被配置為執行本文所描述的方法的各個態樣。

【0158】 儘管在圖10中沒有圖示，但是電子設備1000亦可以包括一或多個資料匯流排，用於在電子設備1000的各個態樣之間傳遞資料。

附加的考慮

【0159】 先前描述被提供用於使任何熟習此項技術者能夠實踐本文描述的各個實施例。本文論述的實例不是對請求項中闡述的範疇、適用性或實施例的限制。對該等實施例的各種修改對熟習此項技術者而言將是顯而易見的，並且本文定義的一般原理可被應用於其他實施例。例如，可以對所論述的元件的功能和佈置進行改變，而不脫離本案的範疇。各種實例可以酌情省略、替代或添加各種程序或元件。例如，可以按與所描述次序不同的次序來執行所描述的方法，以及可以添加、省略，或組合各個步驟。同樣，相對於某些實例描述的特徵可以被組合至某些其他實例中。例如，可以使用任意數量的本文所陳述的態樣來實現裝置或實踐方法。此外，揭示的範疇意欲覆蓋使用作為本文闡述的揭示的各種態樣的附加或除了本文闡述的揭示的各種態樣之外的其他結構、功能性或者結構和功能性來實

踐的裝置或方法。應當理解的是，本文所揭示的揭示的任何態樣可以由請求項中的一或多個元素來體現。

【0160】 如本文所使用的，「示例性」一詞的意思是「用作示例、實例或說明」。本文中被描述為「示例性」的任何態樣並不一定要被解釋為相比其他態樣更佳或有利。

【0161】 如本文所使用的，代表項目列表中的「至少一個」的短語是指彼等項目的任何組合，包括單個成員。作為實例，「a、b或c中的至少一個」意欲覆蓋a、b、c、a-b、a-c、b-c和a-b-c，以及與多個相同元素的任何組合（例如、a-a、a-a-a、a-a-b、a-a-c、a-b-b、a-c-c、b-b、b-b-b、b-b-c、c-c和c-c-c或a、b和c的任何其他次序）。

【0162】 如本文中所使用的，術語「決定」涵蓋各種各樣的動作。例如，「決定」可以包括運算、計算、處理、推導、調查、檢視（例如，在表、資料庫或另一資料結構中檢視）、探明等。同樣，「決定」可以包括接收（例如，接收資訊）、存取（例如，存取記憶體中的資料）等。同樣，「決定」可以包括解析、選擇、選取、建立等。

【0163】 本文揭示的方法包括用於實現該等方法的一或多個步驟或動作。方法步驟及/或動作可以彼此互換而不背離請求項的範疇。換言之，除非步驟或動作的特定次序被指定，否則可以修改特定步驟及/或動作的次序及/或使用而不背離請求項的範疇。此外，上述方法的各種操作可以由能夠執行對應功能的任何合適的構件來執行。該構件可以包括各種硬體及/或軟體元件（一或多個）及/或模組（一

或多個)，包括但不限於電路、特殊應用積體電路（A S I C）或處理器。通常而言，在附圖中圖示操作的情況下，彼等操作可以具有對應的帶有類似編號的配套構件加功能（m e a n s - p l u s - f u n c t i o n）元件。

【0164】 以下請求項不意欲限制於本文所示的實施例，而是符合與請求項的語言一致的全部範疇。在請求項內，除非特別說明，否則以單數形式對元件的引用並不意欲意指「一個且僅一個」，而是「一或多個」。除非另有特定說明，術語「某些」代表一或多個。依據專利法施行細則第18條第8項的規定，將不解釋任何請求項元素，除非使用短語「用於……的構件」來明確陳述元素，或者在方法請求項的情況下使用短語「用於……的步驟」來陳述元素。一般技術者已知或以後將知道的，貫穿本案所描述的各個態樣的元件的全部結構和功能均等物均經由引用被明確地併入本文，並且意欲由請求項涵蓋。而且，本文揭示的內容皆不意欲專用於公眾，無論此類揭示是否在請求項中被明確陳述。

【符號說明】

【0165】

1 0 2 : 增強型記憶體模組

1 0 4 : 記憶體整合加速器

1 0 6 : 快閃記憶體

1 0 8 : 計算核心

1 1 0 A : 記憶體元件

1 1 0 B : 記 憶 體 元 件

1 1 2 A : 記 憶 體 元 件

1 1 2 B : 記 憶 體 元 件

1 1 2 C : 記 憶 體 元 件

1 1 2 D : 記 憶 體 元 件

1 1 2 E : 記 憶 體 元 件

1 1 2 F : 記 憶 體 元 件

1 1 4 A : 鄰 近 記 憶 體 緩 衝 器

1 1 4 B : 鄰 近 記 憶 體 緩 衝 器

1 1 4 C : 鄰 近 記 憶 體 緩 衝 器

1 1 4 D : 鄰 近 記 憶 體 緩 衝 器

1 1 4 E : 鄰 近 記 憶 體 緩 衝 器

1 1 4 F : 鄰 近 記 憶 體 緩 衝 器

1 1 6 : 引 腳

2 0 0 : 處 理 系 統

2 0 2 : O S 使 用 者 空 間

2 0 4 : 模 型

2 0 6 : 資 料

2 0 8 : M L / A I 應 用 軟 體

2 1 0 : 加 速 器 運 行 時 間 應 用 軟 體

2 1 2 : O S 核 心 空 間

2 1 4 : 記 憶 體 驅 動 器

2 1 6 : 加 速 器 驅 動 器

2 1 8 : C P U

2 2 0 : 記 憶 體 管 理 單 元 (M M U)

2 2 2 : 記 憶 體 控 制 器

2 2 4 A : 增 強 型 記 憶 體 模 組

2 2 4 B : 增 強 型 記 憶 體 模 組

2 2 6 A : 驅 動 器

2 2 6 B : 驅 動 器

3 0 2 : 增 強 型 記 憶 體 模 組

3 0 6 : 快 閃 記 憶 體

3 0 8 : 計 算 核 心

3 1 0 A : 專 用 記 憶 體 元 件

3 1 0 B : 專 用 記 憶 體 元 件

3 1 2 A : 記 憶 體 元 件

3 1 2 B : 記 憶 體 元 件

3 1 2 C : 記 憶 體 元 件

3 1 2 D : 記 憶 體 元 件

3 1 2 E : 記 憶 體 元 件

3 1 2 F : 記 憶 體 元 件

3 1 4 A :

3 1 4 B :

3 1 4 C :

3 1 4 D :

3 1 4 E :

3 1 4 F :

4 0 0 : 方 法

- 4 0 2 : 步 驟
- 4 0 4 : 步 驟
- 4 0 6 : 步 驟
- 4 0 8 : 步 驟
- 4 1 0 : 步 驟
- 4 1 2 : 步 驟
- 4 1 4 : 步 驟
- 4 1 6 : 步 驟
- 5 0 0 : 方 法
- 5 0 2 : 步 驟
- 5 0 4 : 步 驟
- 5 0 6 : 步 驟
- 5 0 8 : 步 驟
- 5 1 0 : 步 驟
- 5 1 2 : 步 驟
- 5 1 4 : 步 驟
- 6 0 0 : 方 法
- 6 0 2 : 步 驟
- 6 0 4 : 步 驟
- 6 0 6 : 步 驟
- 6 0 8 : 步 驟
- 6 1 0 : 步 驟
- 6 1 2 : 步 驟
- 6 1 4 : 步 驟

7 0 0 : 方 法

7 0 2 : 步 驟

7 0 4 : 步 驟

7 0 6 : 步 驟

7 0 8 : 步 驟

7 1 0 : 步 驟

7 1 2 : 步 驟

7 1 4 : 步 驟

7 1 6 : 步 驟

7 1 8 : 步 驟

8 0 2 : 記 憶 體 映 射

8 0 4 : 記 憶 體 映 射

8 0 6 : 記 憶 體 映 射

8 0 8 : 記 憶 體 映 射

9 0 0 : 方 法

9 0 2 : 步 驟

9 0 4 : 步 驟

9 0 6 : 步 驟

9 0 8 : 步 驟

9 1 0 : 步 驟

9 1 2 : 步 驟

9 1 4 : 步 驟

1 0 0 0 : 電 子 設 備

1 0 0 2 : C P U

1 0 0 4 : 圖 形 處 理 單 元 (G P U)

1 0 0 6 : 數 位 信 號 處 理 器 (D S P)

1 0 0 8 : 神 經 處 理 單 元 (N P U)

1 0 2 2 : 輸 入 / 輸 出 設 備

1 0 2 4 : 記 憶 體

1 0 2 4 A : 接 收 元 件

1 0 2 4 B : 儲 存 元 件

1 0 2 4 C : 傳 遞 元 件

1 0 2 4 D : 處 理 元 件

1 0 2 4 E : 發 送 (或 提 供) 元 件

1 0 2 4 F : 初 始 化 元 件

1 0 2 4 G : 分 配 元 件

1 0 2 4 H : 分 割 元 件

1 0 2 4 I : 計 算 核 心

1 0 2 4 J : 快 閃 記 憶 體 元 件

【生物材料寄存】

國 內 寄 存 資 訊 (請 依 寄 存 機 構 、 日 期 、 號 碼 順 序 註 記)

無

國 外 寄 存 資 訊 (請 依 寄 存 國 家 、 機 構 、 日 期 、 號 碼 順 序 註 記)

無

【發明申請專利範圍】

【請求項1】 一種增強型記憶體模組，包括：

一計算核心；及

一或多個雙模式記憶體元件，

該一或多個雙模式記憶體元件中的每一雙模式記憶體元件可動態分配至該計算核心以進行記憶體模組上處理，或分配至一主機處理系統，

其中該增強型記憶體模組被配置為：

接收一工作負載的至少一部分，其中該工作負載的該部分由一加速模組根據該增強型記憶體模組的一容量及用於該工作負載的增強型記憶體模組的一總數決定；

根據所接收的該工作負載的該部分及來自一主機處理系統的動態配置資訊，將該一或多個雙模式記憶體元件中的記憶體的一第一子集分配為預留給該主機處理系統使用的主機處理系統可定址記憶體，並且將該一或多個雙模式記憶體元件中的記憶體的一第二子集分配為該記憶體主機處理系統無法使用的計算核心可定址記憶體；

從該主機處理系統接收資料；

利用該計算核心處理該資料以產生經處理的資料；
及

經由記憶體的該第一子集向該主機處理系統提供該經處理的資料。

【請求項2】 如請求項1之增強型記憶體模組，其中該增強型記憶體模組亦被配置為：

將所接收的該資料儲存在記憶體的該第一子集中；

在利用該計算核心處理該資料之前，從記憶體的該第一子集向記憶體的該第二子集傳遞所接收的該資料；及

在利用該計算核心處理該資料之後，從記憶體的該第二子集向記憶體的該第一子集傳遞該經處理的資料。

【請求項3】 如請求項1之增強型記憶體模組，其中記憶體的該第二子集亦包括被配置為供該計算核心使用的一或多個單模式記憶體。

【請求項4】 如請求項1之增強型記憶體模組，亦包括：一快閃記憶體模組，包括用於分配記憶體的該第一子集和記憶體的該第二子集的韌體指令。

【請求項5】 如請求項1之增強型記憶體模組，其中該增強型記憶體模組亦被配置為：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之前，在該主機處理系統的一資料匯流排上設置一忙狀態以指示資料的一傳遞正在發生。

【請求項6】 如請求項5之增強型記憶體模組，其中該增強型記憶體模組亦被配置為：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之後，在該主機處理系統的該資料匯流排上設置一可用狀態。

【請求項7】 如請求項1之增強型記憶體模組，其中該增強型記憶體模組亦被配置為：在從記憶體的該第一子集

向記憶體的該第二子集傳遞該資料之前，啟用對用於該增強型記憶體模組的主機處理系統記憶體命令的一阻止。

【請求項 8】 如請求項 7 之增強型記憶體模組，其中該增強型記憶體模組亦被配置為：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之後，禁用對用於該增強型記憶體模組的主機處理系統記憶體命令的該阻止。

【請求項 9】 如請求項 1 之增強型記憶體模組，其中該資料包括：

用於由一機器學習模型處理的資料；及
機器學習模型參數。

【請求項 10】 如請求項 1 之增強型記憶體模組，其中該增強型記憶體模組包括一雙列直插式記憶體模組（DIMM）。

【請求項 11】 如請求項 1 之增強型記憶體模組，其中該增強型記憶體模組亦被配置為：

解除分配記憶體的該第一子集和記憶體的該第二子集；
及

將該一或多個雙模式記憶體元件中的記憶體的一第三子集分配為主機處理系統可定址記憶體，並且將該一或多個雙模式記憶體元件中的記憶體的一第四子集分配為計算核心可定址記憶體；

其中記憶體的該第三子集不同於記憶體的該第一子集，

以及

其中記憶體該第四子集不同於記憶體該第二子集。

【請求項12】如請求項1所述之增強型記憶體模組，其中執行該記憶體的分配是為了最大化該增強型記憶體模組及該一或多個其他增強型記憶體模組在整個該工作負載及在該增強型記憶體模組和該一或多個其他增強型記憶體模組上執行的該一或多個其他工作負載上的效能。

【請求項13】一種用於利用包括一計算核心的一增強型記憶體模組處理資料的方法，包括以下步驟：

接收一工作負載的至少一部分，其中該工作負載的該部分由一加速模組根據該增強型記憶體模組的一容量及用於該工作負載的增強型記憶體模組的一總數決定；

經由根據所接收的該工作負載的該部分及來自一主機處理系統的動態配置資訊將該增強型記憶體模組的記憶體的一第一子集分配為預留給該主機處理系統使用的主機處理系統可定址記憶體並將該增強型記憶體模組的記憶體的一第二子集分配為該記憶體主機處理系統無法使用的計算核心可定址記憶體，初始化該增強型記憶體模組；

其中記憶體的該第一子集的一大小和記憶體的該第二子集的大小由該主機處理系統動態配置，且

其中記憶體的該第一子集和記憶體的該第二子集與一或多個雙模式記憶體元件相關聯，該一或多個雙模式記

記憶體元件中的每一元件可動態分配至計算核心進行記憶體模組上處理或分配至主機處理系統；

利用該增強型記憶體模組上的計算核心處理資料以生成經處理的資料；及

經由記憶體的該第一子集向主機處理系統提供該經處理的資料。

【請求項 14】如請求項 13 所述之方法，亦包括以下步驟：將所接收的資料儲存在記憶體的該第一子集中；在利用該計算核心處理該資料之前，將該資料從記憶體的該第一子集傳遞至記憶體的該第二子集；及在利用該計算核心處理該資料之後，將該經處理的資料從記憶體的該第二子集傳遞至記憶體的該第一子集。

【請求項 15】如請求項 14 之方法，亦包括以下步驟：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之前，在該主機處理系統的一資料匯流排上設置一忙狀態以指示資料的一傳遞正在發生。

【請求項 16】如請求項 15 之方法，亦包括以下步驟：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之後，在該主機處理系統的該資料匯流排上設置一可用狀態。

【請求項 17】如請求項 14 之方法，亦包括以下步驟：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之前，啟用對用於該增強型記憶體模組的主機處理系

統記憶體命令的一阻止。

【請求項18】如請求項17之方法，亦包括以下步驟：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之後，禁用對用於該增強型記憶體模組的主機處理系統記憶體命令的該阻止。

【請求項19】如請求項13所述之方法，其中記憶體的該第一子集和記憶體的該第二子集與一或多個雙模式記憶體元件相關聯。

【請求項20】如請求項13所述之方法，其中初始化該增強型記憶體模組包括處理儲存在該增強型記憶體模組上的一快閃記憶體模組中的韌體指令。

【請求項21】如請求項13所述之方法，其中記憶體的該第二子集亦與被配置為供該計算核心使用的一或多個單模式記憶體元件相關聯。

【請求項22】如請求項13之方法，其中該資料包括：
用於由一機器學習模型處理的資料；及
機器學習模型參數。

【請求項23】如請求項13之方法，其中該增強型記憶體模組包括一雙列直插式記憶體模組（DIMM）。

【請求項24】如請求項13之方法，亦包括以下步驟：
解除分配記憶體的該第一子集和記憶體的該第二子集；
及

將該增強型記憶體模組的記憶體的一第三子集分配為主機處理系統可定址記憶體，並且將該增強型記憶體模

組的記憶體的一第四子集分配為計算核心可定址記憶體，

其中記憶體的該第三子集不同於記憶體的該第一子集，以及

其中記憶體的該第四子集不同於記憶體的該第二子集。

【請求項25】一種包括指令的非暫時性電腦可讀取媒體，該等指令在由一主機處理系統的一或多個處理器執行時，使得該主機處理系統執行用於利用包括一計算核心的一增強型記憶體模組處理資料的一方法，該方法包括以下步驟：

接收一工作負載的至少一部分，其中該工作負載的該部分由一加速模組根據該增強型記憶體模組的一容量及用於該工作負載的增強型記憶體模組的一總數決定；

經由根據所接收的該工作負載的該部分及來自一主機處理系統的動態配置資訊將該增強型記憶體模組的記憶體的一第一子集分配為預留給該主機處理系統使用的主機處理系統可定址記憶體並將該增強型記憶體模組的記憶體的一第二子集分配為該記憶體主機處理系統無法使用的計算核心可定址記憶體，初始化該增強型記憶體模組；

其中記憶體的該第一子集和記憶體的該第二子集與一或多個雙模式記憶體元件相關聯，該一或多個雙模式記憶體元件中的每一元件可動態分配至計算核心進行記憶

體模組上處理或分配至主機處理系統；

在該增強型記憶體模組處從該主機處理系統接收資料；

利用該增強型記憶體模組上的該計算核心處理該資料以產生經處理的資料；及

經由記憶體的該第一子集向該主機處理系統提供該經處理的資料。

【請求項 26】如請求項 25 之非暫時性電腦可讀取媒體，其中該方法亦包括以下步驟：

將所接收的該資料儲存在記憶體的該第一子集中；

在利用該計算核心處理該資料之前，從記憶體的該第一子集向記憶體的該第二子集傳遞所接收的該資料；及

在利用該計算核心處理該資料之後，從記憶體的該第二子集向記憶體的該第一子集傳遞該經處理的資料。

【請求項 27】如請求項 26 之非暫時性電腦可讀取媒體，其中該方法亦包括以下步驟：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之前，在該主機處理系統的一資料匯流排上設置一忙狀態以指示資料的一傳遞正在發生。

【請求項 28】如請求項 27 之非暫時性電腦可讀取媒體，其中該方法亦包括以下步驟：在從記憶體的該第一子集向記憶體的該第二子集傳遞該資料之後，在該主機處理系統的該資料匯流排上設置一可用狀態以允許。

【請求項 29】如請求項 26 之非暫時性電腦可讀取媒體，

其中該方法亦包括以下步驟：

在從記憶體的第一子集向記憶體的第二子集傳遞該資料之前，啟用對用於該增強型記憶體模組的主機處理系統記憶體命令的一阻止；及

在從記憶體的第一子集向記憶體的第二子集傳遞該資料之後，禁用對用於該增強型記憶體模組的主機處理系統記憶體命令的該阻止。

【請求項30】如請求項25所述之非暫時性電腦可讀取媒體，其中記憶體的第二子集亦與被配置為供該計算核心使用的一或多個單模式記憶體元件相關聯。

【發明圖式】

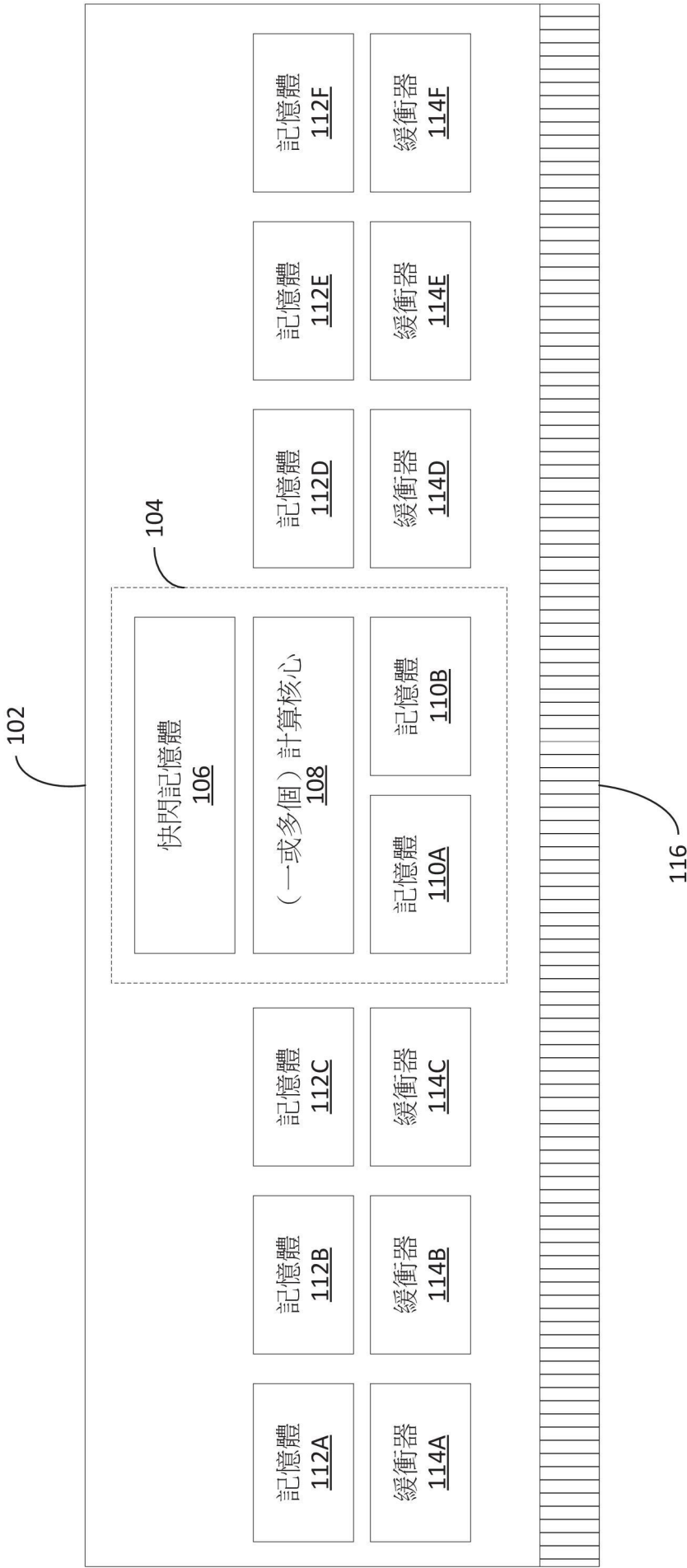


圖1

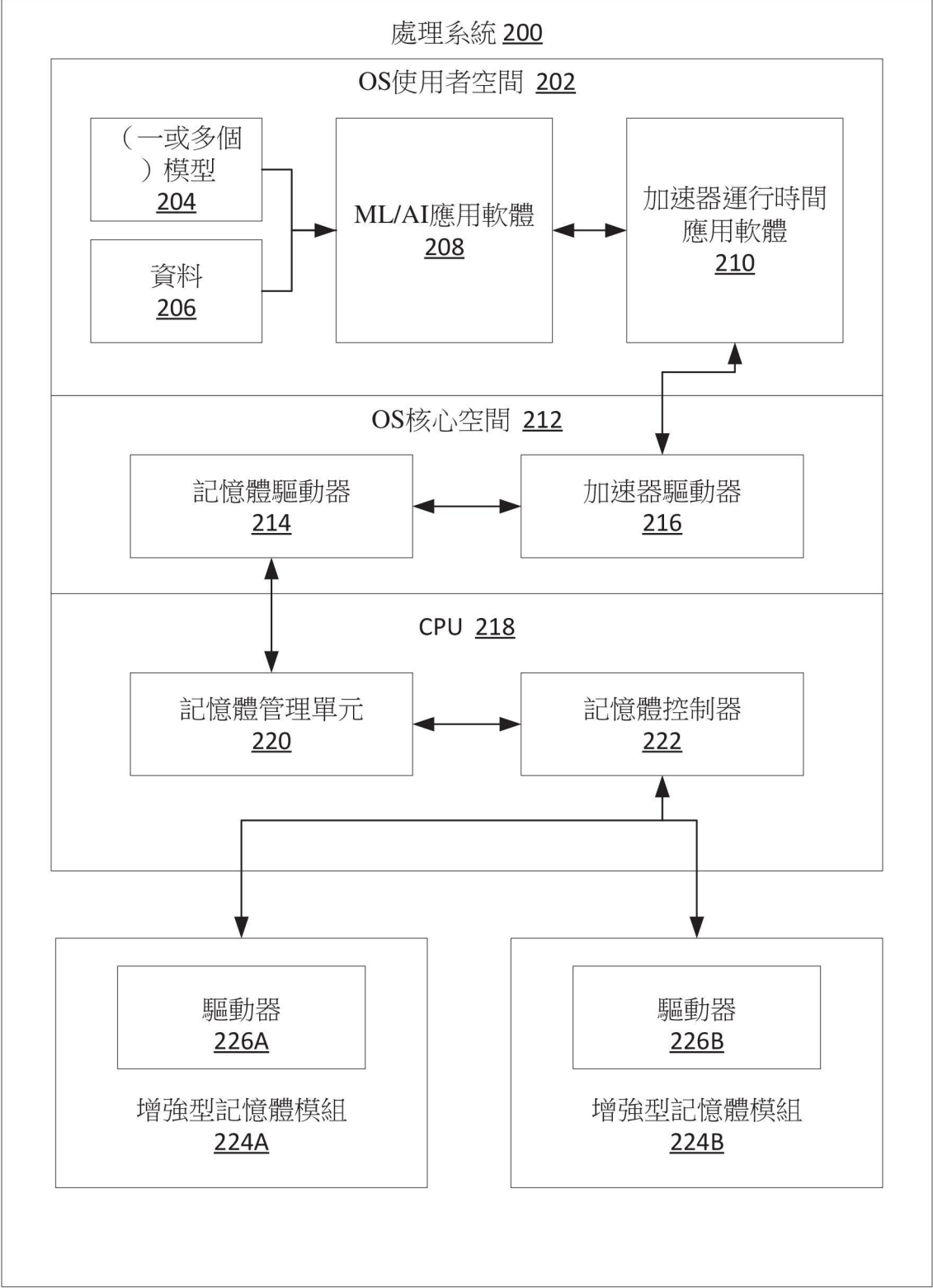


圖2

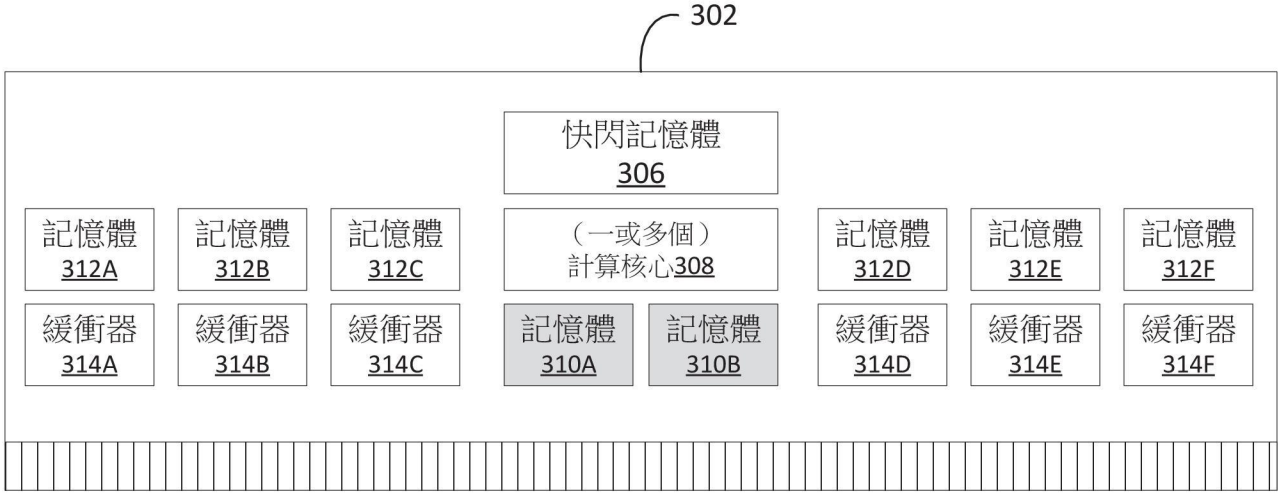


圖3A

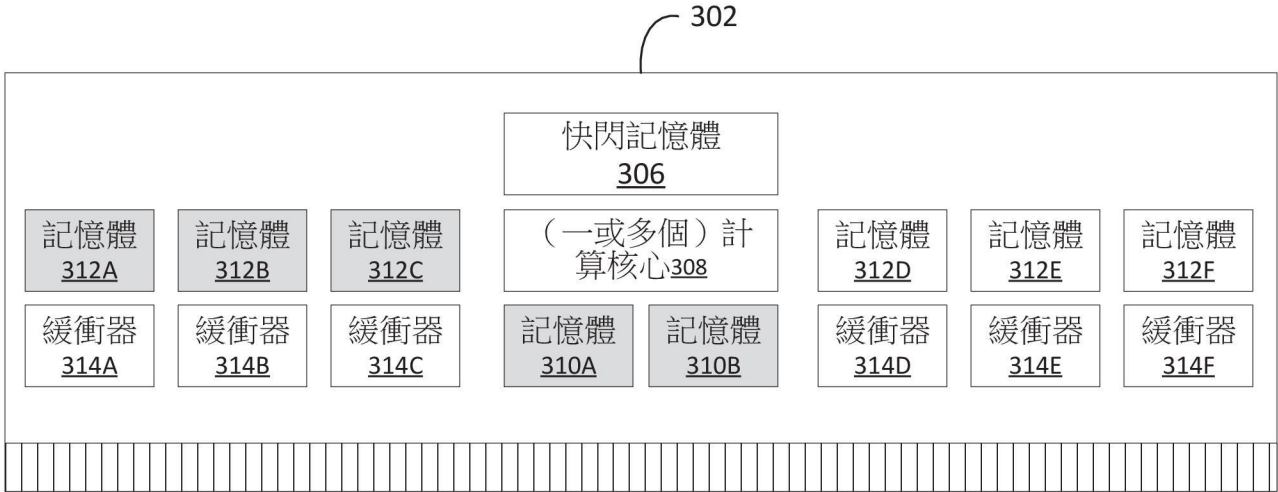


圖3B

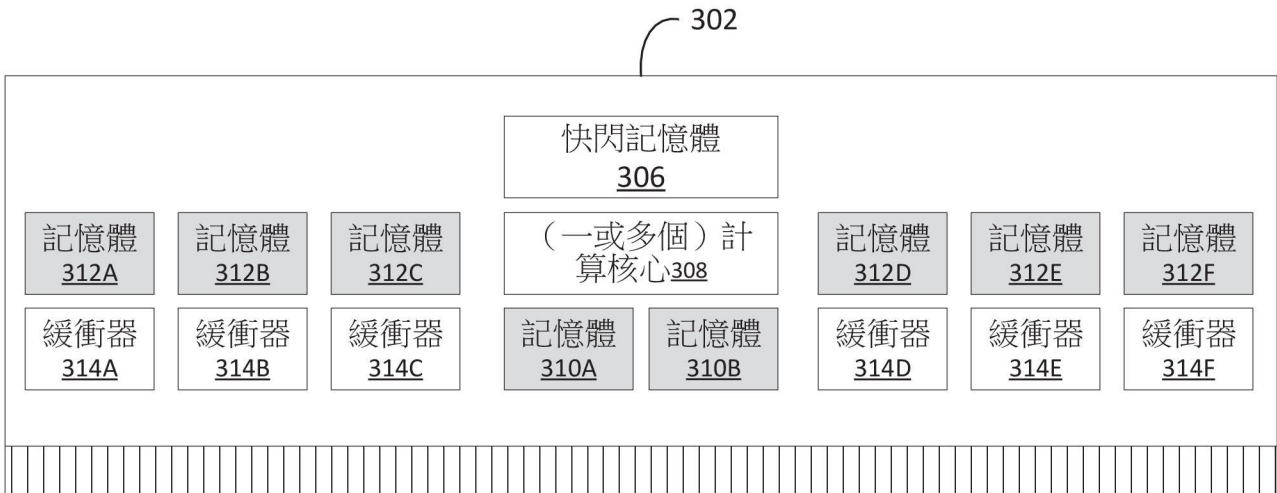


圖3C

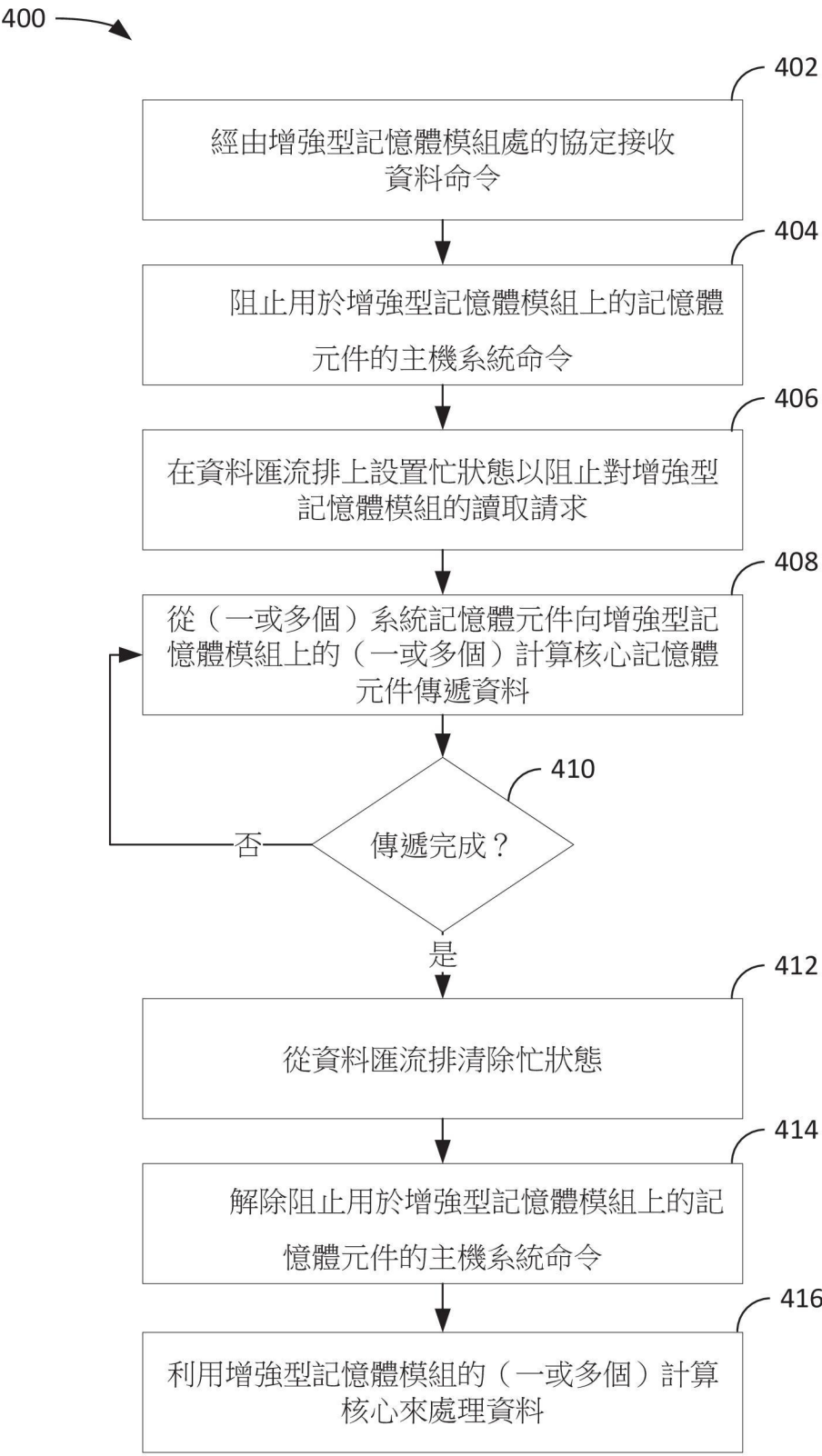


圖4

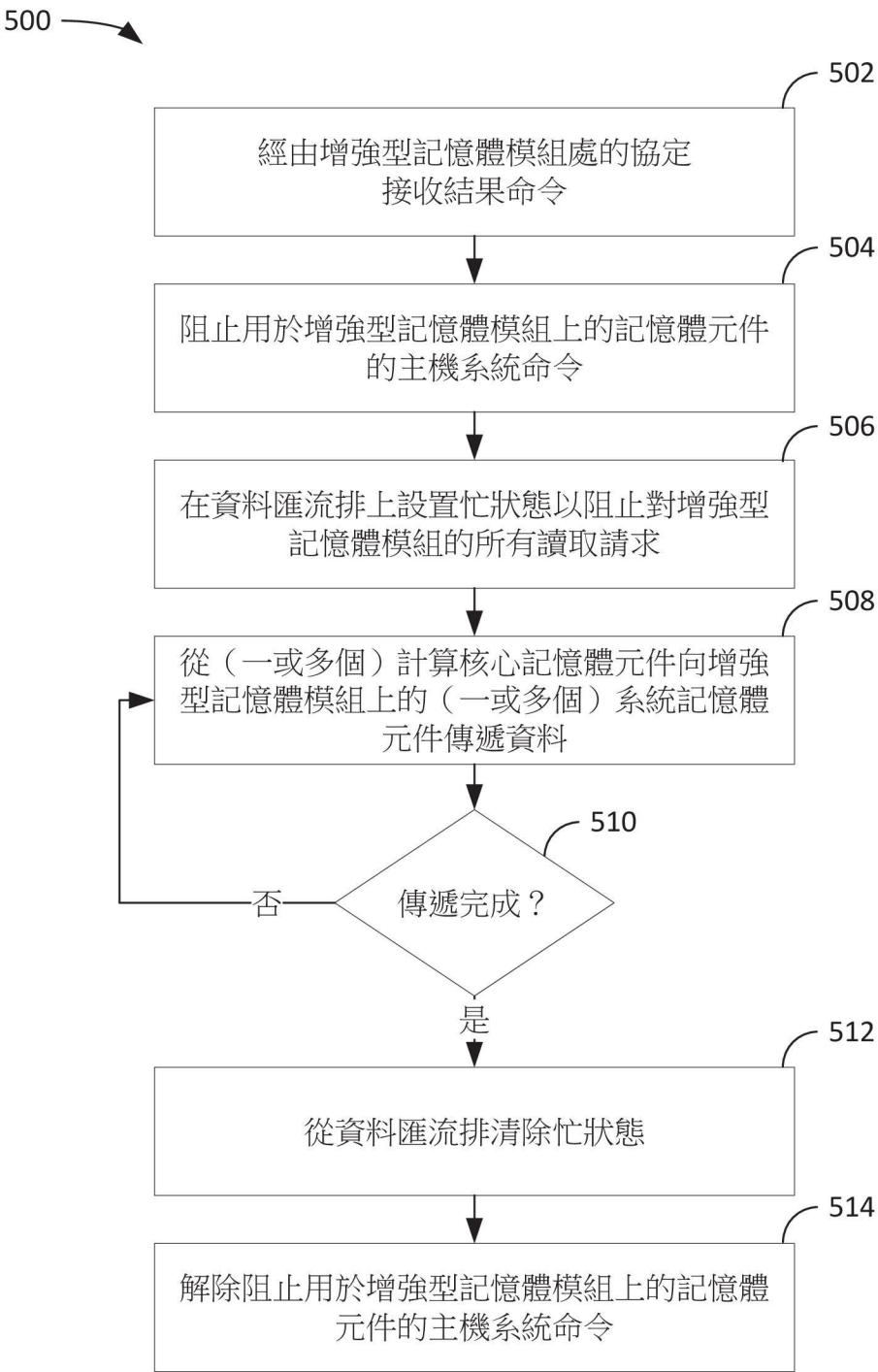


圖5

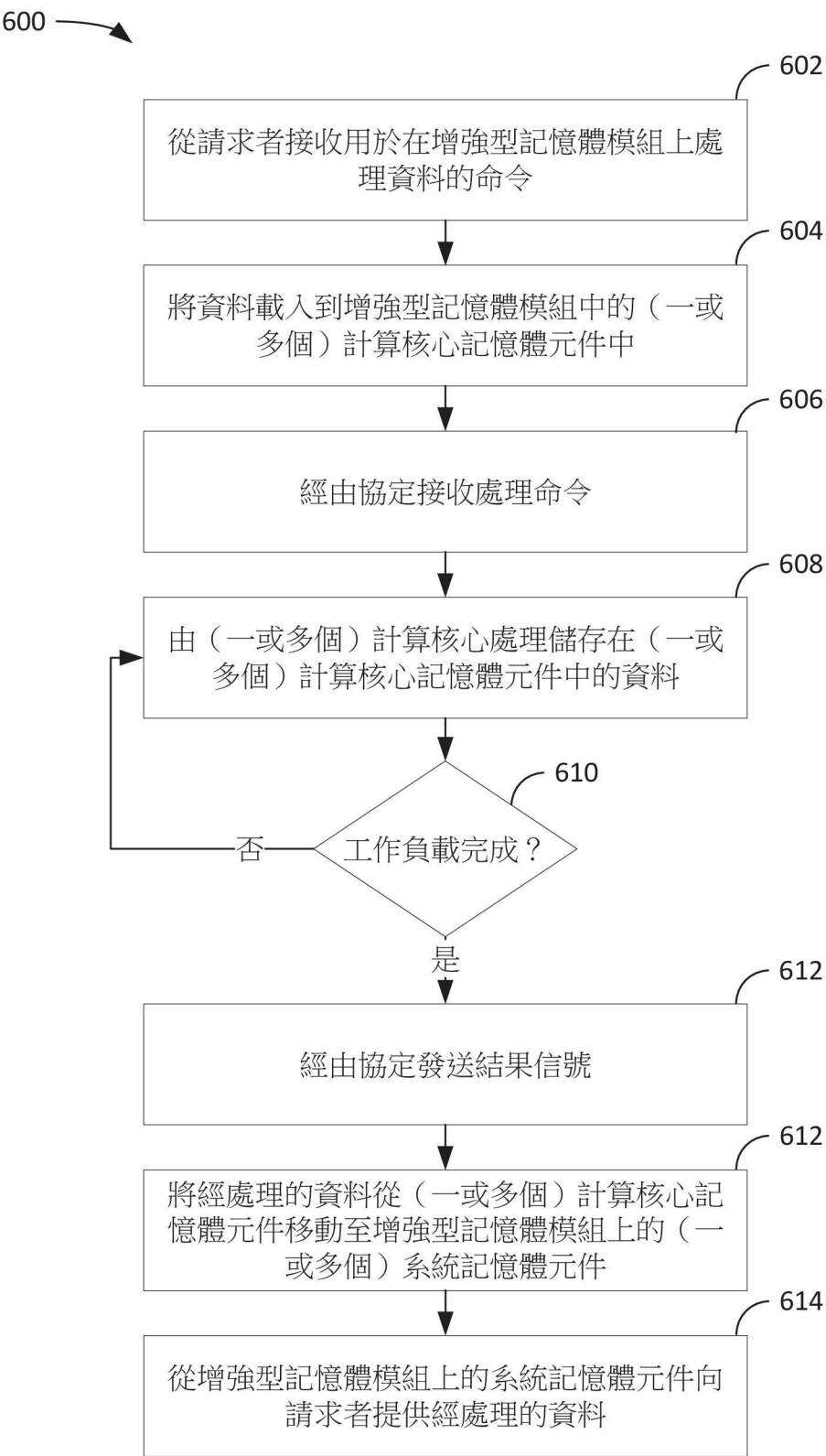


圖6

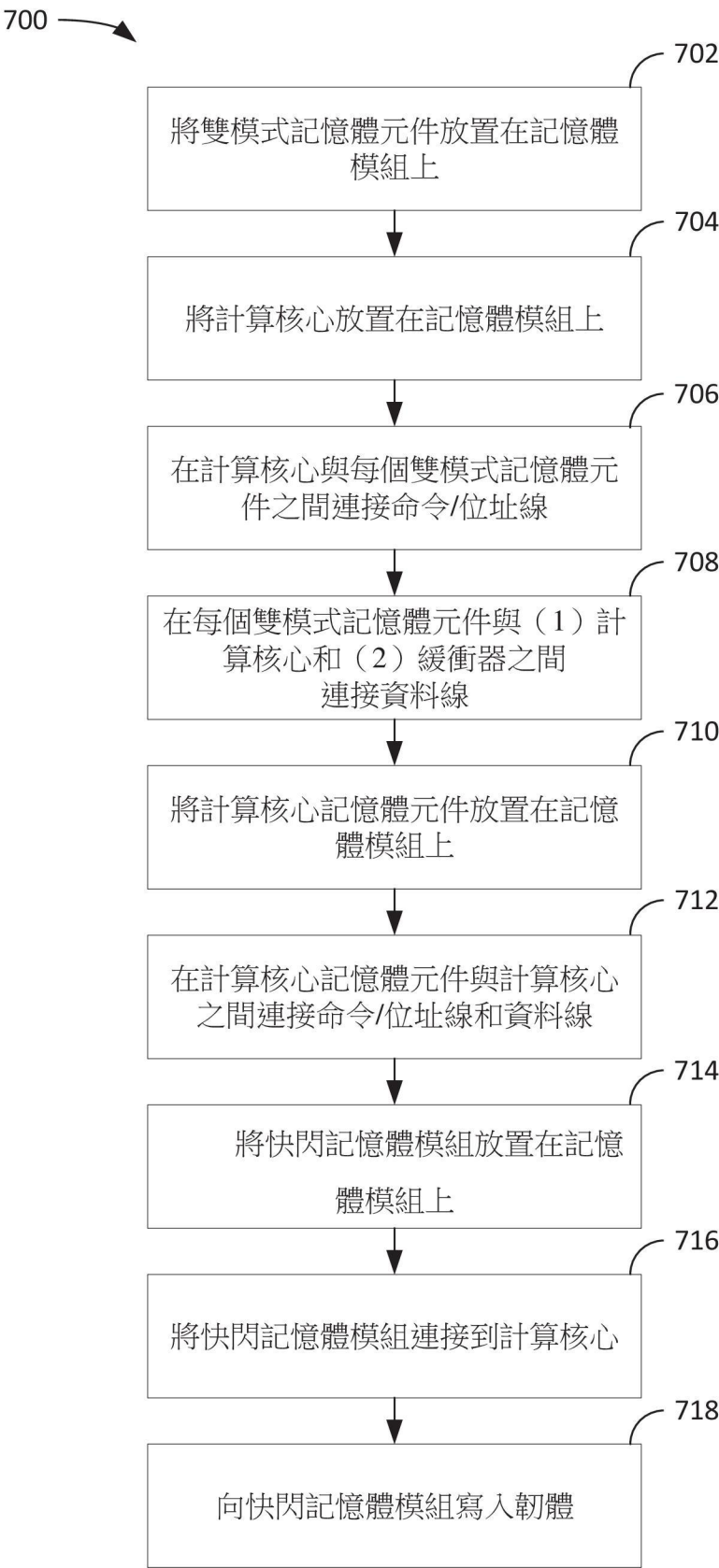


圖7

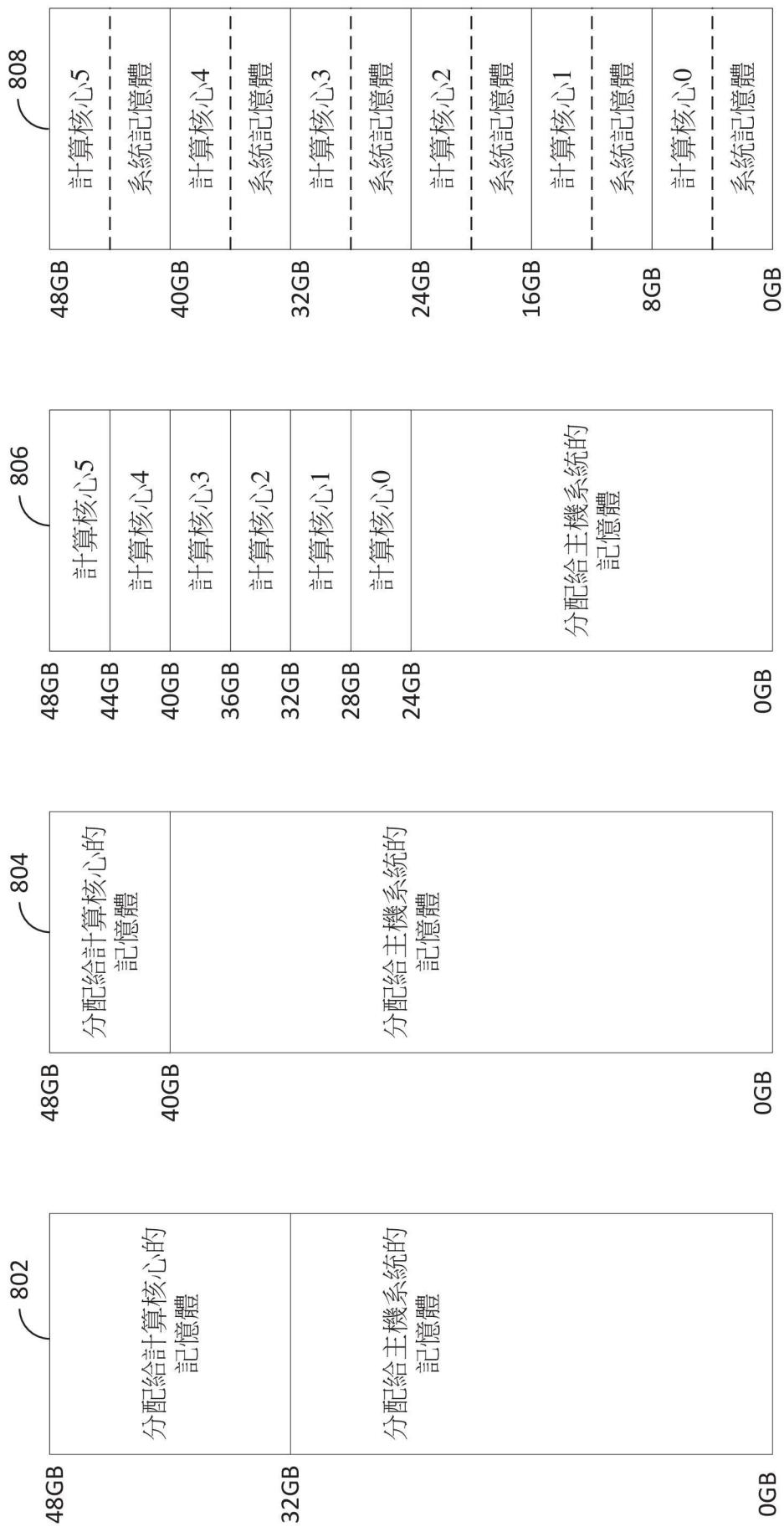


圖8

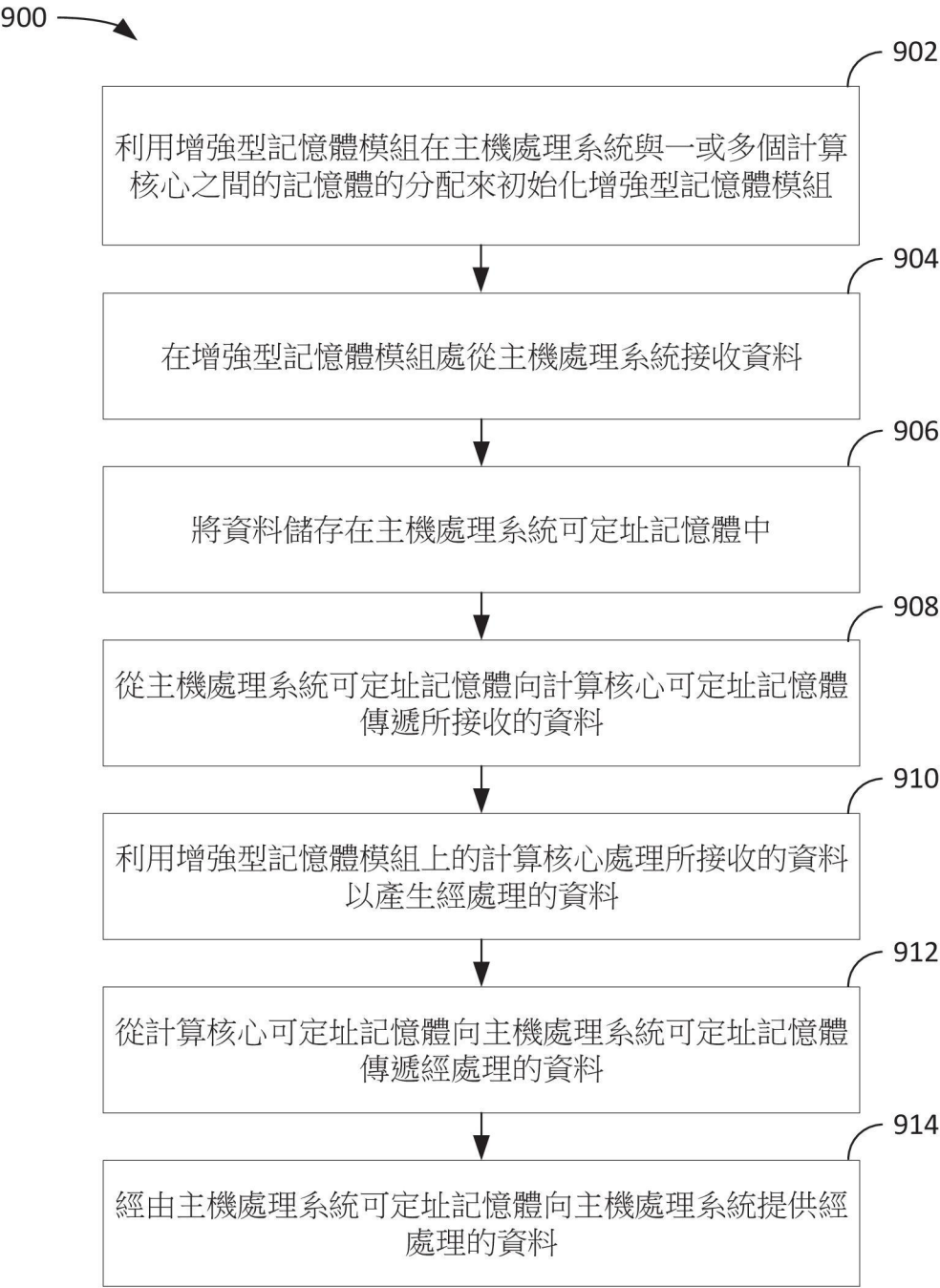


圖9

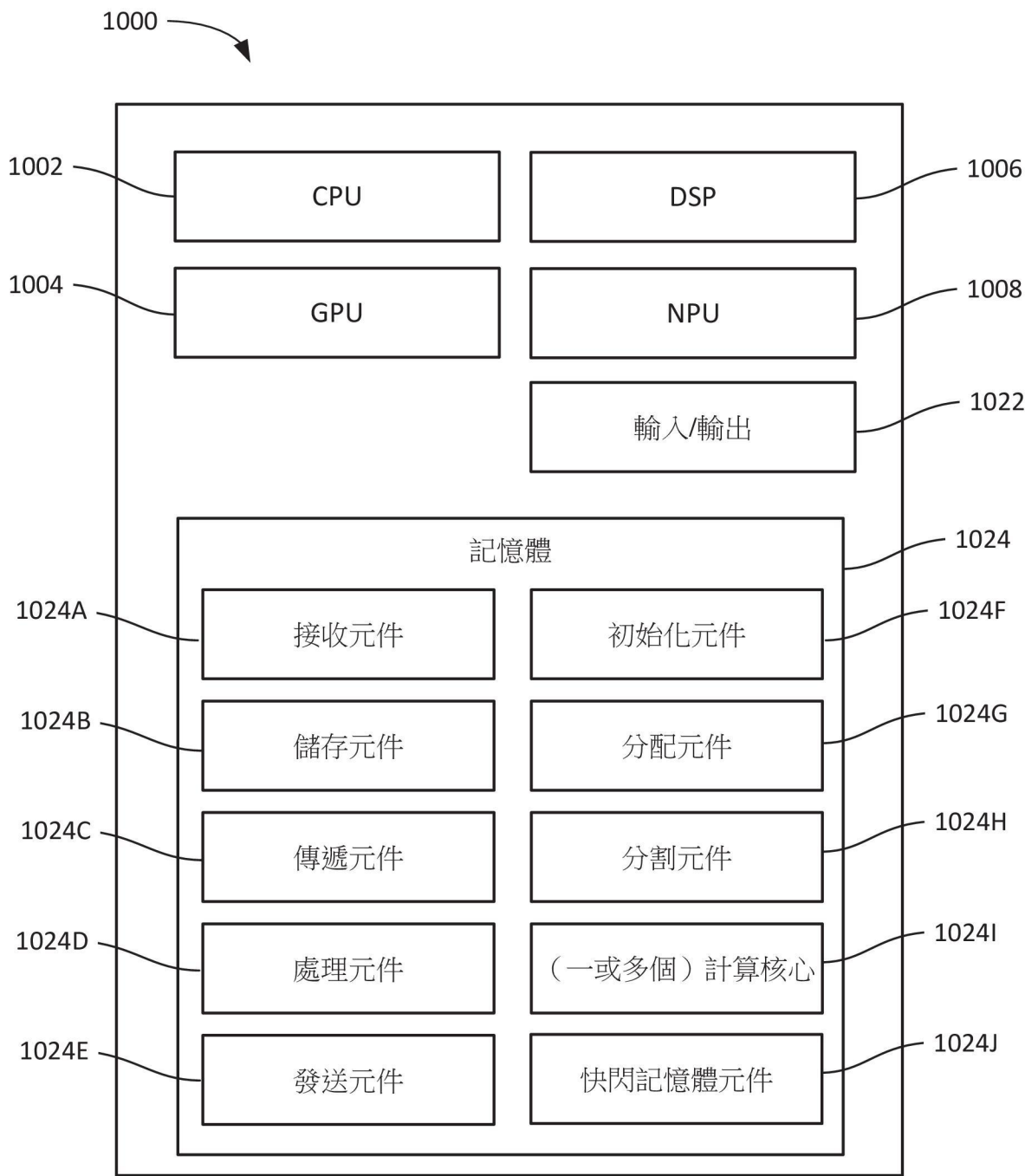


圖10