I816566

(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 I816566 B**

(45)公告日： 中華民國 112 (2023) 年 09 月 21 日

(21)申請案號：111136960 (22)申請日： 中華民國 108 (2019) 年 11 月 27 日

(51)Int. Cl.： ***G06F12/02 (2006.01)*** ***G06F13/10 (2006.01)***
G06F13/14 (2006.01)

(30)優先權：2019/04/29 美國 16/397,481

(71)申請人：美商谷歌有限責任公司 (美國) GOOGLE LLC (US)
美國

(72)發明人：瑪達爾 勞倫斯 J 三世 MADAR, LAWRENCE J., III (US)；菲德魯 泰米達尤 FADELU, TEMITAYO (NG)；凱坦 哈爾希 KHAITAN, HARSHIT (IN)；納拉亞那斯瓦密 拉非 NARAYANASWAMI, RAVI (IN)

(74)代理人：陳長文

(56)參考文獻：

TW 201911039A US 2018/0322390A1

審查人員：簡大翔

申請專利範圍項數：20 項 圖式數：3 共 38 頁

(54)名稱

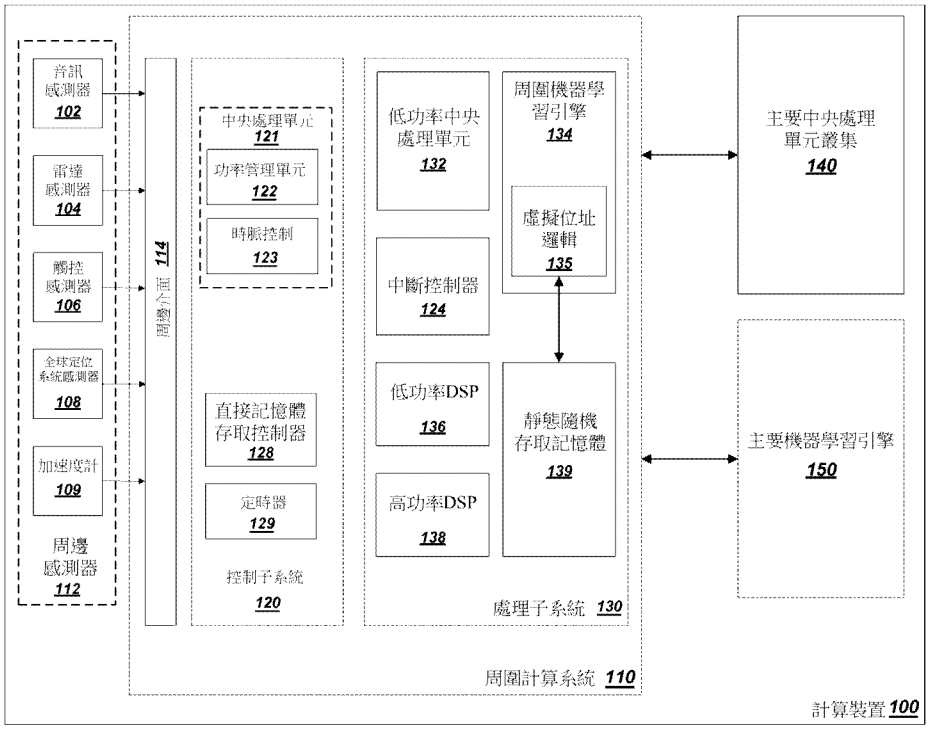
用於將外部記憶體虛擬化為機器學習加速器之局部記憶體之裝置及系統

(57)摘要

本發明揭示用於將外部記憶體虛擬化為一機器學習加速器之局部記憶體之方法、系統及設備，包含編碼於電腦儲存媒體上之電腦程式。一種周圍計算系統包括：一周圍機器學習引擎；一低功率CPU；及一 SRAM，其至少在該周圍機器學習引擎及該低功率 CPU 當中共用；其中該周圍機器學習引擎包括虛擬位址邏輯以自由該周圍機器學習引擎產生之虛擬位址變換為該 SRAM 內之實體位址。

Methods, systems, and apparatus, including computer programs encoded on computer storage media, for virtualizing external memory as local to a machine learning accelerator. One ambient computing system comprises: an ambient machine learning engine; a low-power CPU; and an SRAM that is shared among at least the ambient machine learning engine and the low-power CPU; wherein the ambient machine learning engine comprises virtual address logic to translate from virtual addresses generated by the ambient machine learning engine to physical addresses within the SRAM.

指定代表圖：



【圖1】

符號簡單說明：

100:實例性計算裝置/裝置/計算裝置

102:音訊感測器

104:雷達感測器

106:觸控感測器

108:全球定位系統感測器

109:加速度計

110:周圍計算系統

112:感測器/周邊感測器

114:周邊介面

120:控制子系統

121:功率控制單元

122:功率管理單元

123:時脈控制單元

124:中斷控制器

128:直接記憶體存取控制器

129:定時器

130:處理子系統

132:低功率中央處理單元

134:周圍機器學習引擎

135:虛擬位址邏輯

136:低功率 DSP

138:高功率 DSP

139:靜態隨機存取記憶體/共用靜態隨機存取記憶體

140:主要中央處理單元叢集

150:主要機器學習引擎

I816566

【發明摘要】

【中文發明名稱】

用於將外部記憶體虛擬化為機器學習加速器之局部記憶體之裝置及系統

【英文發明名稱】

DEVICE AND SYSTEM FOR VIRTUALIZING EXTERNAL MEMORY AS LOCAL TO A MACHINE LEARNING ACCELERATOR

【中文】

本發明揭示用於將外部記憶體虛擬化為一機器學習加速器之局部記憶體之方法、系統及設備，包含編碼於電腦儲存媒體上之電腦程式。一種周圍計算系統包括：一周圍機器學習引擎；一低功率CPU；及一SRAM，其至少在該周圍機器學習引擎及該低功率CPU當中共用；其中該周圍機器學習引擎包括虛擬位址邏輯以自由該周圍機器學習引擎產生之虛擬位址變換為該SRAM內之實體位址。

【英文】

Methods, systems, and apparatus, including computer programs encoded on computer storage media, for virtualizing external memory as local to a machine learning accelerator. One ambient computing system comprises: an ambient machine learning engine; a low-power CPU; and an SRAM that is shared among at least the ambient machine learning engine and the low-power CPU; wherein the ambient machine learning engine comprises virtual address logic to translate from virtual addresses generated by the ambient machine learning engine to physical addresses within the SRAM.

【指定代表圖】

圖1

【代表圖之符號簡單說明】

100:實例性計算裝置/裝置/計算裝置

102:音訊感測器

104:雷達感測器

106:觸控感測器

108:全球定位系統感測器

109:加速度計

110:周圍計算系統

112:感測器/周邊感測器

114:周邊介面

120:控制子系統

121:功率控制單元

122:功率管理單元

123:時脈控制單元

124:中斷控制器

128:直接記憶體存取控制器

129:定時器

130:處理子系統

132:低功率中央處理單元

134:周圍機器學習引擎

135:虛擬位址邏輯

136:低功率DSP

138:高功率DSP

139:靜態隨機存取記憶體/共用靜態隨機存取記憶體

140:主要中央處理單元叢集

150:主要機器學習引擎

【發明說明書】

【中文發明名稱】

用於將外部記憶體虛擬化為機器學習加速器之局部記憶體之裝置及系統

【英文發明名稱】

DEVICE AND SYSTEM FOR VIRTUALIZING EXTERNAL MEMORY AS LOCAL TO A MACHINE LEARNING ACCELERATOR

【技術領域】

【先前技術】

【0001】 本說明書係關於機器學習加速器。

【0002】 一機器學習(「ML」)加速器係一裝置或一裝置上之一組件，例如，具有經設計以用於有效地訓練機器學習模型、執行機器學習模型或既訓練機器模型又執行機器學習模型之一專業架構的一積體電路。

【0003】 一ML加速器可經組態以透過一或多個機器學習模型執行推理遍次。每一推理遍次使用輸入及一機器學習模型之習得參數值來產生由習得模型預測之一或多個輸出。該ML加速器可包含一或多個計算塊。一般而言，一計算塊係經組態以獨立地執行一組計算之一自足式計算組件。一ML加速器之塊可配置成一網路且經程式化使得該ML加速器之每一塊經組態以透過機器學習模型執行一推理遍次之一個部分之操作。舉例而言，若該機器學習模型係一神經網路，則主要ML引擎150中之每一塊可經組態以計算該神經網路之一個層之計算。

【0004】 ML加速器需要大量記憶體來靈活地處理不同種類之機器學習模型。若一ML加速器係一周圍計算裝置(例如，一手機或其他計算裝置)中之一組件(其可處於一低功率狀態中，仍監測來自環境之輸入且對該

等輸入做出回應)，則此要求產生至少兩個問題：

【0005】 首先，分配一上部記憶體範圍以適應較大機器學習模型對於使用一ML加速器之大多數周圍計算裝置係成本太高的。另外，此方法亦導致專用於一ML加速器之記憶體浪費，該ML加速器通常可執行僅需要經分配記憶體之一分率之模型。實體空間亦浪費在考慮到可移植性而設計之一周圍計算裝置之一緊密包裝晶片上。分配給該ML加速器之記憶體亦不可用於裝置之其他處理組件。

【0006】 其次，未用於處理機器學習模型之記憶體仍自裝置汲取電力，從而浪費能量。此問題在考慮到低功率消耗而設計之諸多周圍計算裝置上惡化。

【發明內容】

【0007】 本說明書闡述用於虛擬化外部記憶體以供一ML加速器使用之技術。該ML加速器可包含(例如)實施為一積體電路之邏輯以用於變換由該ML加速器存取之虛擬記憶體位址同時處理或訓練一機器學習模型。該等虛擬記憶體位址變換為在ML加速器外部(諸如RAM上或以通信方式連接至在其中實施ML加速器之處理子系統之一系統級快取記憶體上)之記憶體位置。具有對應參數之機器學習模型可自ML加速器外側串流式傳輸且由ML加速器存取以模擬對在ML加速器之局部記憶體中之位置之讀取及寫入。

【0008】 本說明書中所闡述之標的物之特定實施例可經實施以便實現以下優點中之一或多者。一小的低功率ML加速器可經實施以處理由一周圍計算裝置接收之周圍信號。該ML加速器可係存取較大共用快取記憶體而不將較大快取記憶體之存取僅限定於ML加速器之一單個計算塊。

【0009】 虛擬記憶體位址可經指派以經濟地串流式傳輸機器學習模

型及對應參數，此乃因ML加速器可使一虛擬記憶體位址範圍擴展或收縮，如由一特定機器學習模型之記憶體要求所指示。該ML加速器可存取一機器學習模型及儲存於在該ML加速器外部之記憶體上之參數而不必須重新串流式傳輸資料，例如，不必須在一卷積神經網路中重新串流式傳輸通常重用之模型參數。此外，該ML加速器可存取在不具有任何特殊組態之情況下儲存於外部之機器學習模型及參數。換言之，自該ML加速器之視角來看，該ML加速器似乎正在存取在加速器之局部記憶體。

【0010】 類似地，在具有彼記憶體組態之裝置中，該ML加速器亦可存取儲存於一系統級快取記憶體上之資料。分配虛擬記憶體位址之能力避免對專用於ML加速器之大量記憶體之需要，藉此減少功率消耗及ML加速器必須在周圍計算裝置中佔據之實體空間。

【0011】 經編譯以用於對ML加速器執行之機器學習模型可使用在不具有廣泛組態或定製化之情況下揭示之記憶體虛擬化特徵。用於一機器學習模型之一編譯器可藉助指示應使用記憶體虛擬化之最少額外指令來編譯模型。

【0012】 在附圖及下文之說明中陳述本說明書之標的物之一或多個實施例之細節。根據說明、圖式及申請專利範圍將明瞭標的物之其他特徵、態樣及優點。

【圖式簡單說明】

【0013】

圖1係實施具有虛擬位址邏輯之一ML加速器之一實例性計算裝置之一圖式。

圖2係實施以通信方式連接至圖1之計算裝置之一SLC的一SOC上之一實例性系統之一圖式。

圖3係用於使用記憶體虛擬化對一ML加速器執行一機器學習模型之一實例性程序之一流程圖。

在各個圖式中，相似元件符號及名稱指示相似元件。

【實施方式】

【0014】 本說明書闡述用於實施將記憶體虛擬化為一機器學習(「ML」)加速器之局部記憶體之技術。替代將大量記憶體專用為ML加速器之局部記憶體，ML加速器可透過由在ML加速器局部之虛擬位址邏輯自動指派之虛擬記憶體位址存取外部記憶體。用於ML加速器之虛擬位址邏輯亦可包含邏輯以取決於在經由模型實施推理遍次之一程式之編譯期間添加至一機器學習模型之一組態選項而接通或關斷此記憶體虛擬化特徵。

【0015】 在將一機器學習模型編譯為可由一ML加速器(諸如下文所論述之一周圍ML引擎134或一主要ML引擎150)執行之經編譯指令期間可包含此及其他組態選項。

【0016】 圖1係實施具有虛擬位址邏輯之一ML加速器之一實例性計算裝置100之一圖式。裝置100可包含在任何適當計算裝置(例如，一智慧型電話、一智慧型手錶、一健身追蹤器、一個人數位助理、一電子平板電腦、一膝上型電腦，僅舉幾個例子)中實施之一周圍計算系統110。可使用計算裝置100之周圍計算系統110，使得計算裝置100可保持在一低功率狀態中，仍藉由順序地喚醒系統之適當處理組件而不斷地監測來自環境之輸入且對該等輸入做出回應。雖然在此處關於圖1論述周圍計算系統110，但對實施低功率周圍計算系統之一大體論述可存在於以其全文引用方式併入本文中之第PCT/US2018/062329號國際申請案中。

【0017】 計算裝置100之一或多個組件可實施於計算裝置內之一系統單晶片(「SOC」)上。一SOC可係一積體電路，該積體電路在一單個矽

基板上或在(例如)使用矽插入器、堆疊式晶粒或互連橋接件之多個互連晶粒上包含系統之每一組件。計算裝置之其他組件(包含一主要CPU叢集140、一主要ML引擎150或一處理子系統130)可實施於同一晶粒上或一單獨晶粒上。

【0018】 計算裝置100可包含與SOC分開且獨立於SOC之組件(包含感測器112、一或多個顯示器、一電池及其他組件)，且可(舉例而言)安裝於一共同殼體上。計算裝置100包含用於控制將電力及感測器信號供應至系統中之組件之一控制子系統120。裝置100包含用於處理感測器信號且產生輸出之一處理子系統130。

【0019】 裝置100可包含若干個周邊感測器112。周邊感測器112可包含一或多個音訊感測器102、一或多個雷達感測器104、一或多個觸控感測器106、一全球定位系統(「GPS」)感測器108及/或一加速度計109。系統可包含額外、較少或替代周邊感測器。周邊感測器112可係經組態以回應於環境輸入而產生感測器信號之裝置。

【0020】 周圍計算系統110可包含一或多個周邊介面114。周邊介面114可係甚至在裝置處於其最低功率狀態中時接通電源的計算裝置100之一組件。周邊介面114可包含用於將自周邊感測器112接收之輸入轉換為待由周圍計算系統110使用之感測器信號之任何適當周邊介面。

【0021】 周邊介面114中之每一者經組態以在偵測到一環境輸入之後即刻產生一各別中斷。一般而言，每一中斷可識別感測器資料之一源，例如，負責中斷之一周邊介面或感測器之一識別符。該等中斷由一或多個中斷控制器124接收且處理。舉例而言，在接收到一中斷之後，中斷控制器124即刻可喚醒一功率控制單元(「PCU」) 121，功率控制單元121包含一功率管理單元(「PMU」) 122及一時脈控制單元123。PMU 122可控制

裝置100之哪些組件接收電力及每一組件接收多少電力。時脈控制單元123可控制裝置100之組件操作之頻率。

【0022】 在本說明書中，無論何時將感測器信號闡述為至其他處理組件之輸入，該等輸入皆可係由感測器自身產生之類比電信號、感測器信號之數位表示或表示原始信號之一或多個性質的感測器信號之經處理數位表示。

【0023】 在接收到一中斷之後，PCU 121即刻可基於中斷之源而判定應啟動周圍計算系統110之哪些其他組件以便進一步處理導致中斷之感測器信號。為了為此等組件提供處理支援，PCU 121可喚醒靜態隨機存取記憶體(「SRAM」) 139及系統通信組構。

【0024】 系統通信組構係以通信方式耦合周圍計算系統110之內部組件、其與外部組件之通信或此等各項之某一組合的一通信子系統。該組構可包含通信硬體(例如，匯流排或專用互連電路系統)之任何適當組合。

【0025】 儘管未繪示，但計算裝置100亦可包含通常存在於此等計算裝置上之一或多個其他組件，例如，一顯示器、一數據機、一圖形處理單元、一顯示器處理器或一特殊用途影像處理器，僅舉幾個例子。此等組件可在下文所闡述之低功率狀態期間關斷電源且在系統判定感測器信號匹配需要其啟動之一應用程式之情況下啟動。

【0026】 裝置100亦包含一主要CPU叢集140。主要CPU叢集140係可包含與處理子系統130中之組件分開之一或多個一般用途處理器的計算裝置100之一組件。主要CPU叢集140之處理器一般具有比處理子系統130中之組件中之任一者多之計算功率，且因此，主要CPU叢集140之處理器亦可消耗比處理子系統130中之組件中之任一者多之電力。

【0027】 控制子系統120亦可包含一定時器129，定時器129係可偵

測系統故障且解決彼等故障之一電子定時器。在正常操作期間，控制子系統120可定期重啟定時器129以阻止定時器129逾時。若控制子系統120 (例如)由於一硬體故障或一程式錯誤而未能重啟一定時器，則該定時器將時間流逝且產生一逾時信號。該逾時信號可用於起始一或多個校正動作。一校正動作可包含將周圍計算系統110放置於一安全狀態中且恢復正常系統操作。

【0028】 處理子系統130包含一周圍機器學習引擎134。周圍ML引擎134係經組態以透過一或多個機器學習模型執行推理遍次之一特殊用途處理裝置。

【0029】 周圍ML引擎134可包含用於計算神經網路啟動或其他神經網路層輸出之一或多個乘法累加(「MAC」)單元及一或多個總和暫存器，以及用於控制總和暫存器與MAC單元之間的資料交換之一控制器。周圍ML引擎134亦可包含指令記憶體、直接記憶體存取路徑、暫存器及其他處理組件。在某些實施方案中，周圍ML引擎134係經組態以使機器學習推理遍次之計算加速之一單個機器學習計算塊。

【0030】 周圍ML引擎包含虛擬位址邏輯135。虛擬位址邏輯135可係可將由周圍ML引擎產生之虛擬位址變換為SRAM 139中之實體記憶體位址的周圍ML引擎134中之一專門電路，SRAM 139係用於周圍ML引擎134之非局部記憶體。在本說明書中，用於計算裝置100之一組件(例如，周圍ML引擎134)之非局部記憶體係指由該組件及一或多個其他組件使用之記憶體。換言之，非局部記憶體未由該組件排他地使用。

【0031】 舉例而言，SRAM 139可係可由處理子系統130之多個處理組件(例如，低功率DSP 136、高功率DSP 138、低功率CPU 132以及周圍ML引擎134)共用之一般用途靜態隨機存取記憶體裝置。因此，SRAM係

用於周圍ML引擎134之非局部記憶體。相比之下，周圍ML引擎134亦可包含由周圍ML引擎134排他地使用且可整合至與周圍ML引擎之其餘部分相同之矽晶粒中之局部記憶體。舉例而言，周圍ML引擎134可具有包含一或多個整合式暫存器之局部記憶體。該等整合式暫存器係用於周圍ML引擎134之局部記憶體，此乃因暫存器中之資料可僅自周圍ML引擎134讀取或僅由周圍ML引擎134寫入。

【0032】 SRAM 139可儲存感測器信號、處理器指令與資料、系統輸出及其他資料，例如，由周圍ML引擎134實施或將由周圍ML引擎134實施之神經網路模型之神經網路參數。

【0033】 一般而言，一SRAM與DRAM之區別在於：SRAM不需要週期性地再新。如下文更詳細地闡述，處理子系統130中之處理組件可直接或透過直接記憶體存取(「DMA」)控制器128存取SRAM 139。在某些實施方案中，SRAM 139包含多個記憶體組，每一記憶體組具有實質上類似資料容量，例如，各自具有1 MB、10 MB或100 MB。另外，每一個別記憶體組可包含可在進入一低功率狀態時個別地斷開電源之多個記憶體區塊。藉由仔細地對區塊在多個記憶體組當中斷開電源之次序進行定序，SRAM記憶體位址空間可保持連續的。

【0034】 虛擬位址邏輯135可變換周圍ML引擎134產生之虛擬記憶體位址。在某些實施方案中，虛擬位址邏輯135 (例如)使用由周圍ML引擎134產生之虛擬位址之最高有效位維持虛擬頁與實體頁之間的一映射。

【0035】 虛擬位址邏輯135可接收自由周圍ML引擎134所執行之經編譯指令產生之一虛擬位址讀取或對該虛擬位址進行寫入之一請求。虛擬位址邏輯135然後可將虛擬位址映射至SRAM 139中之一實體位址。在某些實施方案中，虛擬位址邏輯135將一虛擬頁碼映射至一實體頁碼且複製

最低有效位以產生實體位址。

【0036】 將一虛擬記憶體位址變換為一實體記憶體位址意味：替代地，當一ML加速器執行用於一機器學習模型之經編譯指令時，對經編譯指令中之一記憶體位址中之資料之每個讀取或寫入指令致使一對應實體記憶體位址位置處之資料經讀取或寫入。在某些實施方案中，虛擬位址邏輯135經組態以回應於周圍ML引擎134執行對映射至對應實體記憶體位址之虛擬記憶體位址位置處之資料之一讀取或寫入指令而將讀取或寫入指令發佈至SRAM 139中之一對應實體記憶體位址位置。

【0037】 虛擬位址邏輯135可將虛擬記憶體位址映射至並非在周圍ML引擎134局部之記憶體中之位置之實體記憶體位址。在某些實施方案中，處理子系統130可在由經映射實體記憶體位址引用之位置處重寫現有資料。虛擬位址邏輯135可經組態以自動地或回應於由對機器學習模型執行推理遍次之一經編譯程式指示之一組態選項(例如，作為待對周圍ML引擎134執行之在該經編譯程式上之一或多個指令)而執行此初始化步驟。

【0038】 周圍ML引擎134可藉由存取經分配非局部記憶體而執行具有使用一機器學習模型執行一推理遍次之一或多個指令之一經編譯程式。自周圍ML引擎134之視點來看，當實際上自一共用記憶體源(諸如自SRAM 139或一系統級快取記憶體)存取資料時，將在虛擬記憶體位址位置處存取之資料視為在周圍ML引擎134局部。

【0039】 裝置100亦可視情況包含一主要ML引擎150。主要ML引擎150係經組態以透過一或多個機器學習模型執行推理遍次(即，對主要ML引擎150執行機器學習模型)之一特殊用途處理裝置。與周圍ML引擎134一樣，每一推理遍次使用輸入及一機器學習模型之習得參數值來產生由習得模型預測之一或多個輸出。主要ML引擎150可包含一或多個計算塊，該

一或多個計算塊可配置成一網路且經程式化使得主要ML引擎150之每一塊經組態以透過機器學習模型執行一推理遍次之一個部分之操作。在以其全文引用方式併入本文中之第9,710,265號美國專利中闡述具有多個計算塊之一適合機器學習引擎。

【0040】 當裝置100包含一主要ML引擎150及一周圍ML引擎134兩者時，周圍ML引擎134一般具有較少計算塊且因此具有比主要ML引擎150少之處理功率且消耗比主要ML引擎150少之電力。舉例而言，周圍ML引擎134可實施為一個或兩個計算塊，然而主要ML引擎150可具有8個或更多個互連塊。

【0041】 每一計算塊可具有在塊局部之小量之記憶體。在計算塊局部之記憶體量通常不足以用於單獨處理一機器學習模型，這就是為什麼(如上文所論述)個別塊可組態成一網路以共用資源且分配針對一給定輸入處理一機器學習模型之任務作為指派給每一計算塊之一系列子任務。

【0042】 由於周圍ML引擎134一般以相對少於主要ML引擎150之計算塊來實施，因此可用之一個或兩個計算塊可不足以用於處理一機器學習模型，甚至在最佳化一網路組態或在塊當中分割處理之後。此可因為計算塊可缺乏處理機器學習模型之計算能力，或因為計算塊可不具有充足記憶體。在某些實施方案中，甚至最基本網路化或分割係不可獲得的，此乃因周圍ML引擎134實施為一單個計算塊。因此，虛擬位址邏輯135可執行如周圍ML引擎134執行一機器學習模型所需要之記憶體虛擬化。

【0043】 儘管圖1中未展示，但主要ML引擎150亦可包含虛擬位址邏輯以虛擬化用於主要ML引擎150之非局部記憶體。雖然主要ML引擎150一般具有比周圍ML引擎134多之計算資源，但主要ML引擎150亦可需要存取不在主要ML引擎150局部之記憶體以執行特定機器學習模型。在

彼等情形中，可使用針對周圍ML引擎134中之虛擬位址邏輯135所闡述之相同技術來實施用於主要ML引擎150之虛擬位址邏輯。替代將虛擬記憶體位址變換為如關於周圍ML引擎134所闡述之SRAM中之實體記憶體位址，用於一主要ML引擎150之虛擬位址邏輯可經組態以將虛擬記憶體位址變換為一系統級快取記憶體(「SLC」)之實體記憶體位址。

【0044】 一SLC可係可快取自記憶體擷取之資料或待儲存於一系統中之多個不同硬體裝置之記憶體中之資料的一裝置或一裝置之一組件，例如，計算裝置100。換言之，SLC之不同快取線可儲存屬不同硬體裝置之資料。在某些實施方案中且如下文關於圖2所論述，虛擬位址邏輯可在主要ML引擎150上實施以變換對應於SLC上之實體記憶體位置之虛擬位址。

【0045】 接下來，呈現對周圍計算系統110之一實例性操作之一論述。在本說明書中，術語「喚醒」及「啟動」將用於意指將經增加量之電力供應至用於電子器件之一特定處理組件或其他電路系統。周圍計算系統110可已經或可尚未將電力供應至經喚醒或啟動之一處理組件或其他電路系統。換言之，經喚醒或啟動之一組件先前可已經或可尚未完全關斷電源。喚醒或啟動一處理組件可使得處理組件執行一開機啟動程序且致使用於處理組件之指令及資料載入至隨機存取記憶體中。另一選擇係或另外，喚醒或啟動一處理組件可包含自一先前暫停狀態重新開始。

【0046】 當PCU 121喚醒SRAM 139時，PCU 121可喚醒SRAM 139之並非所有區塊或所有記憶體組。PCU 121可替代地喚醒足以使處理子系統130之下一組件判定是否進一步使裝置100之組件之加電升級的僅若干個區塊。

【0047】 PCU 121亦可將不同功率級供應至SRAM 139之不同區塊。舉例而言，在監測功率狀態中，PMU 122可將一較低保持電壓供應至

整個SRAM 139以減少其功率消耗。若處理組件不需要存取SRAM 139，則PMU 122亦可將保持電壓供應至SRAM 139。在處理功率狀態中，PMU 122可將正常電壓提供至SRAM 139之全部或部分且將降低電壓或未將電壓提供至SRAM 139之其他部分。

【0048】 在處置一中斷之程序期間，周圍計算系統110亦可喚醒一或多個DMA控制器128。DMA控制器128可管理允許用於傳入感測器信號之較高資料頻寬之DMA路徑。舉例而言，一DMA控制器可用於將音訊資料自一麥克風連續地串流式傳輸至SRAM 139中以由處理子系統130中之處理組件存取。相反地，一DMA控制器亦可用於連續地串流式傳輸儲存於SRAM 139中之音訊資料以透過一或多個揚聲器作為聲音而輸出。DMA控制器128亦可用於將任何適當感測器資料串流式傳輸至SRAM 139中，但使用經程式化IO可比針對小量資料啟動一DMA控制器計算上便宜。因此，周圍計算系統110可針對相對高頻寬感測器資料(例如，音訊資料及雷達資料)啟動且使用DMA控制器128，且可針對其他類型之感測器資料使用經程式化IO。

【0049】 在製備組構及SRAM 139之後，PCU 121然後可使用中斷來判定喚醒處理子系統130之哪些其他組件。舉例而言，PMU 122可取決於一或多個感測器中之哪一者產生一中斷而控制是否將電力提供至低功率CPU 132、低功率DSP 136或處理子系統130之其他組件。在某些實施方案中，周邊介面114及控制子系統120之組件係在一監測功率狀態中接通電源的裝置100之僅有組件，該監測功率狀態係其中周圍計算系統110由於至計算裝置之環境輸入而正在等待接收中斷之一功率狀態。

【0050】 處理子系統130之處理組件可包含一低功率CPU 132、周圍ML引擎134、一低功率DSP 136及一高功率DSP 138。在某些實施方案

中，處理子系統具有此等組件中之一或多者之多個例項，例如，多個低功率DSP或多個高功率DSP。舉例而言，處理子系統130可具有專用於處理音訊信號之一個高功率DSP及專用於處理雷達信號之一單獨高功率DSP。另一選擇係或另外，處理子系統130可具有專用於處理影像資料之一高功率DSP。

【0051】 在監測功率狀態中，處理子系統130中之處理組件可維持在一保持模式中。PCU 121可藉由如下方式將一組件維持於保持模式中：減少或消除提供至該組件之電力。舉例而言，在保持模式中，PCU 121可以剛好足以維持暫存器狀態之電力而非足以處理暫存器中之資料之電力供應一處理組件。

【0052】 低功率CPU 132可係包含暫存器、控制電路系統及一算術邏輯單元(「ALU」)之一般用途可程式化處理器。一般而言，低功率CPU 132消耗比計算裝置之主要CPU叢集140少之電力，且可含有較少處理核心。在某些實施方案中，低功率CPU 132主要係對單個指令及單個資料輸入進行操作之一純量處理器。基於感測器信號之類型，低功率CPU 132接收且基於彼等感測器信號之性質，低功率CPU 132可判定應啟動系統之其他組件，例如，通信組構、DMA控制器128、SRAM 139之某些或全部或此等各項之某一組合。在啟動此等組件之後，低功率CPU 132可視情況返回至一非操作狀態。

【0053】 低功率CPU 132可將感測器信號或其一經處理版本提供至周圍ML引擎134以用於進一步解譯。舉例而言，若低功率CPU 132接收到對應於加速度計輸入之感測器信號，則低功率CPU 132可判定周圍ML引擎134應進一步處理感測器信號。周圍ML引擎134然後可進一步處理感測器信號。

【0054】 周圍ML引擎134之一個任務係使用感測器信號來經由一機器學習模型執行一推理遍次以產生可觸發喚醒其他處理組件以進一步處理感測器信號的一輸出。換言之，周圍ML引擎134可接收由低功率CPU 132或另一處理組件產生之感測器信號或其一經處理版本，且周圍ML引擎134可產生表示哪些其他處理組件應進一步處理感測器信號之一輸出。

【0055】 周圍ML引擎134亦可針對各種不同任務執行機器學習模型，包含針對：計算裝置之一使用者進行之晶片上自動語音辨識、文字至語音產生或手勢辨識。周圍ML引擎134可將因執行一機器學習模型而產生之輸出提供至低功率CPU 132或另一處理組件，以用於額外動作。

【0056】 如上文所論述，虛擬位址邏輯135可經組態以在被指示時(例如，依據用於一機器學習模型之經編譯指令中之一指令)執行記憶體虛擬化。可在由一適當地經組態編譯器編譯機器學習模型期間設定此選項，例如，作為一預設選項或回應於來自用於實施編譯器之一程式之一使用者提示之輸入。用於對周圍ML引擎134執行機器學習模型之經編譯指令保持不變，無論選擇還是不選擇記憶體虛擬化，但編譯器可另外例如藉由一或多個指令指示周圍ML引擎134應使用記憶體虛擬化執行機器學習模型。

【0057】 若啟用記憶體虛擬化，則周圍ML引擎134可使用由虛擬位址邏輯135產生之經映射虛擬記憶體位址自不在周圍ML引擎134局部之記憶體源串流式傳入模型參數及其他模型組態資源。舉例而言，用於機器學習模型之模型參數可儲存於SRAM 139中且由一或多個實體記憶體位址引用。

【0058】 用於機器學習模型之經編譯指令可包含用於載入、讀取且寫入記憶體中之資料之指令。虛擬位址邏輯可經組態使得在一經編譯機器

學習模型之經編譯指令中引用之虛擬記憶體位址變換為對應實體記憶體位址而不更改經編譯指令之記憶體中之參考。

【0059】 作為一額外步驟，處理子系統130可最初將用於機器學習模型之模型參數及其他組態資源自另一記憶體裝置載入至SRAM 139中。舉例而言，DMA控制器128可將模型參數自DRAM串流式傳輸至SRAM 139中。DRAM可在周圍計算系統110局部或外部。然後，虛擬位址邏輯135可將虛擬記憶體位址映射至其中載入有模型參數之SRAM 139中之實體記憶體位置。作為將參數串流式傳輸至SRAM 139中之一部分，DMA控制器128可經組態以重寫儲存於SRAM 139中之現有資料，或另一選擇係將模型參數串流式傳輸至SRAM 139中之可用空間中。機器學習模型之經編譯指令中之一或多個指令可規定DMA控制器128是否應重寫SRAM 139中之現有資料。

【0060】 如上文所論述，由於SRAM 139可包含取決於計算裝置100之狀態而可經啟動或可未經啟動之多個記憶體組，因此SRAM 139中之某些記憶體區塊或組可並非可用的，此乃因存在已經現有資料，或此乃因彼等特定記憶體區塊或組尚未經啟動。

【0061】 雖然DRAM可係周圍計算系統110之一部分，但DRAM亦可在周圍計算系統110外部。在某些實施方案中，DRAM在周圍計算系統110外部，但仍在同一SOC上。在某些實施方案中，DRAM在上面實施有周圍計算系統之SOC外部。在任一實施方案中，DMA控制器128可經組態以自DRAM串流式傳輸模型參數。

【0062】 若模型參數無法串流式傳輸至SRAM 139中，例如，此乃因SRAM 139當前正由周圍計算系統110之其他處理組件使用，則虛擬位址邏輯135可直接將虛擬記憶體位址變換為其中儲存有模型參數之實體記

憶體位址。在某些實施方案中，虛擬位址邏輯可經組態以變換其中儲存有模型參數之DRAM上之實體記憶體位址，而非首先使用DMA控制器128將模型參數串流式傳輸至SRAM 139中。

【0063】 用於機器學習模型之組態資源亦可包含在編譯時間規定之一或多個指令，其指示周圍ML引擎134應存取多少局部記憶體。舉例而言，若編譯器在執行模型之ML引擎將使一特定大小之記憶體分配給其之假定下將一機器學習模型編譯成一組指令，則虛擬位址邏輯可經組態以提供彼大小之記憶體作為映射至SRAM 139中之實體記憶體位址之虛擬記憶體位址。另一選擇係，在編譯時間處，執行編譯器之一程式可回應於一使用者提示或預設條件而設定一記憶體大小。

【0064】 周圍ML引擎134可結合存取儲存在周圍ML引擎134局部之資料而引用如上文所闡述之虛擬記憶體位址。在其中周圍ML引擎134包含一或多個暫存器之某些實施方案中，周圍ML引擎134可透過一虛擬記憶體位址存取不在周圍ML引擎134局部之資料，且然後將經存取資料載入至周圍ML引擎134之暫存器。這樣，最有可能由周圍ML引擎134重複地存取之資料(諸如用於一卷積神經網路模型中之一卷積之參數或共同參數之一子集)可儲存於可用於周圍ML引擎134之最快速記憶體(其通常為局部暫存器)上。

【0065】 由周圍ML引擎134產生之輸出可明確地規定處理組件ID之一組合或一所列舉功率狀態之一識別符，或該輸出可係由一低功率處理組件解譯之一功率狀態之一表示(例如，低功率CPU或低功率DSP)，以便識別應處理感測器信號之其他較高功率處理組件。作為此程序之一部分，低功率處理組件可明確地或含蓄地判定是否需要任何其他處理。舉例而言，低功率處理組件可基於周圍ML引擎134之輸出而判定不需要額外處理且

周圍計算系統110可轉變回至監測功率狀態。

【0066】 在最低級監測功率狀態中，PCU 121可使周圍ML引擎134保持在一低功率狀態中或完全關斷電源。在處理功率狀態中，PCU 121取決於可在周邊介面114處獲得什麼感測器信號及低功率CPU 132或低功率DSP 136如何解譯信號而可或可不將電力提供至周圍ML引擎134。在某些實施方案中，低功率DSP 136或低功率CPU 132可解譯信號以指示PCU 121針對一額外中間功率狀態提供電力，其中周圍ML引擎134亦針對推理遍次接通電源，但其他高功率處理組件尚未接通電源。

【0067】 低功率DSP 136及高功率DSP 138係經組態以用於高度向量化信號之有效解碼及處理之特殊用途處理器。處理子系統130可包含出於不同目的而設計之各種DSP。舉例而言，處理子系統130可包含經組態以處理雷達信號之一DSP或經組態以處理音訊信號之一DSP。

【0068】 如上文所闡述，低功率DSP 136可執行對來自控制子系統120之感測器信號之初始解譯。同樣地，低功率DSP 136亦可執行其他信號處理任務。一般而言，高功率DSP消耗比低功率DSP高之功率級，此乃因其具有更多主動暫存器，其並行地存取且處理更多資料，此乃因其更嚴重地依賴於記憶體操作，或此等各項之某一組合。

【0069】 圖2係實施以通信方式連接至圖1之計算裝置100之一SLC的一SOC 200上之一實例性系統之一圖式。舉例而言，SOC 200可裝設於計算裝置100上或整合至計算裝置100中，或係一單獨裝置或一單獨裝置之組件。

【0070】 計算裝置組件250可透過一SOC組構240與一SLC 230通信。計算裝置組件250可係經組態以能夠與SLC 230通信的計算裝置100上之任何組件，且可包含主要ML引擎150、主要CPU叢集140及周圍計算系

統110。

【0071】 SOC組構240係SOC 200之一通信子系統且可包含允許計算裝置組件250彼此通信而且發出讀取及寫入SLC 230上之資料之請求的通信路徑。SLC 230具有可使用專用暫存器或高速RAM實施之專用快取記憶體。SOC組構240可包含通信硬體(例如，匯流排或專用互連電路系統)之任何適當組合。

【0072】 SOC 200亦包含允許SLC 230與一記憶體控制器220之間的通信之通信路徑252以及允許記憶體控制器220與不在SOC 200局部之DRAM 210之間的通信之晶片間通信路徑254。記憶體控制器220可處置自SLC 230及DRAM 210讀取記憶體且將記憶體寫入至SLC 230及DRAM 210之請求。儘管僅在圖2中展示DRAM 210，但記憶體控制器220可與未展示之其他記憶體裝置(諸如任何揮發性或非揮發性記憶體裝置，例如，一硬碟機或一固態磁碟機)通信。

【0073】 SLC 230可自計算裝置組件250快取讀取請求、寫入請求或兩者。SLC 230可藉由以儲存於快取記憶體中之資料對請求做出回應而非自DRAM 210提取資料而自用戶端裝置快取讀取請求。類似地，SLC可藉由將新資料寫入於快取記憶體中而非將新資料寫入於DRAM中而自用戶端裝置快取寫入請求。SLC 230然後可在一稍後時間執行一回寫以將經更新資料儲存於DRAM 210中。

【0074】 如上文所論述，主要ML引擎150可包含虛擬位址邏輯以將虛擬記憶體位址變換為映射至不在主要ML引擎150局部之記憶體之實體記憶體位址。鑒於虛擬位址邏輯135將實體記憶體位址映射至處理子系統130之共用SRAM 139，在某些實施方案中，用於主要ML引擎150之虛擬位址邏輯可將虛擬記憶體位址映射至位於SLC 230中之實體記憶體位址。

【0075】 當針對對主要ML引擎150執行之一機器學習模型啟用記憶體虛擬化時，虛擬位址邏輯可經組態以透過SOC組構240與SLC 230通信以將虛擬記憶體位址變換為DRAM 210中之實體記憶體位址。作為一初始化步驟，記憶體控制器220可串流式傳入第一次自DRAM 210或不在SOC 200局部之其他記憶體裝置串流式傳輸至SLC 230中之模型參數。

【0076】 圖3係用於使用記憶體虛擬化對一ML加速器執行一機器學習模型之一實例性程序之一流程圖。為了方便，圖3中之程序將闡述為由位於一或多個位置中之一或多個電腦之一系統執行。舉例而言，根據本說明書適當地程式化之一周圍計算系統(例如，圖1之周圍計算系統110)可執行圖3中之程序。用於實施可執行圖3中之程序之一系統之額外細節可存在於上文對圖1及圖2之說明中。

【0077】 系統自不在系統局部之一記憶體裝置串流式傳入模型參數且將該等模型參數串流式傳輸至一共用記憶體裝置(310)中。如上文參考圖1及圖2所論述，用於對一ML加速器執行之一機器學習模型之模型參數可儲存於不在系統局部之記憶體中，諸如DRAM上。舉例而言透過一DMA控制器，系統可串流式傳入用於執行機器學習模型之模型參數及任何組態選項。舉例而言，記憶體裝置可係共用SRAM。在某些實施方案中，如較早論述，系統可重寫共用記憶體裝置中之現有資料。如上文所論述，ML加速器可係系統之周圍ML引擎或主要ML引擎。若ML加速器係系統之主要ML引擎，則共用記憶體裝置可係一系統級快取記憶體。

【0078】 系統透過ML加速器上之虛擬位址邏輯產生虛擬記憶體位址，該等虛擬記憶體位址映射至共用記憶體裝置(320)中儲存模型參數之位置之對應實體記憶體位址。如上文參考圖1所論述，虛擬位址邏輯可依據用於機器學習模型之經編譯指令中之讀取或寫入指令中所引用之記憶體

位址產生虛擬記憶體位址。虛擬記憶體邏輯亦可經組態以產生虛擬記憶體位址與實體記憶體位址之間的映射且將映射儲存於在ML加速器之局部記憶體中，諸如暫存器中。

【0079】 系統藉由執行由虛擬位址邏輯(300)變換之經編譯指令而執行一經編譯程式以對機器學習模型執行一推理遍次。如上文參考圖1所論述，系統執行一經編譯程式，該經編譯程式使用機器學習模型(具體而言使用儲存於共用記憶體裝置中之模型參數)執行一推理遍次。如上文亦論述，虛擬記憶體位址與實體記憶體位址之間的變換意味將適當讀取或寫入指令發佈至一實體記憶體位址位置，該實體記憶體位址位置使經編譯指令中之讀取或寫入指令匹配至映射至該實體記憶體位址之一虛擬記憶體位址之一位置。

【0080】 本說明書中所闡述之標的物及功能性操作之實施例可實施於包含本說明書中所揭示之結構及其結構等效物之數位電子電路系統、有形地體現之電腦軟體或韌體、電腦硬體中，或包含於該等各項中之一或多者之組合中。亦可將本說明書中所闡述之標的物之實施例實施為一或多個電腦程式，亦即，編碼於一有形非暫時性儲存媒體上以供資料處理設備執行或用以控制資料處理設備之操作之一或多個電腦程式指令模組。該電腦儲存媒體可係一機器可讀儲存裝置、一機器可讀儲存基板、一隨機或串列存取記憶體裝置或其中之一或多者之一組合。另一選擇係或另外，程式指令可編碼於經產生以編碼用於傳輸至適合接收器設備以供一資料處理設備執行之資訊的一人工產生之所傳播信號(例如，一機器產生之電、光學或電磁信號)上。

【0081】 術語「資料處理設備」係指資料處理硬體且囊括用於處理資料之所有種類之設備、裝置及機器，藉由實例方式，包含一可程式化處

理器、一電腦或者多個處理器或電腦。該設備亦可係或進一步包含特殊用途邏輯電路系統，例如，一FPGA (場可程式化閘陣列)或一ASIC (特殊應用積體電路)。除硬體以外，該設備亦可視情況包含為電腦程式創建一執行環境之程式碼，例如，構成處理器韌體、一協定堆棧、一資料庫管理系統、一作業系統或其中之一或多者之一組合的程式碼。

【0082】 亦可稱為或闡述為一程式、軟體、一軟體應用程式、一應用程式、一模組、一軟體模組、一指令碼或程式碼之一電腦程式可以任何形式之程式化語言(包含經編譯或經解譯語言，或宣告式或程序式語言)來撰寫，且其可以任何形式來部署，包含作為一獨立程式或作為一模組、組件、副常式或適合於在一計算環境中使用之其他單元。一程式可但不需要對應於一檔案系統中之一檔案。一程式可儲存於保存其他程式或資料(例如，儲存於一標記語言文件中之一或多個指令碼)之一檔案之一部分中、儲存於專用於所討論之程式之單個檔案中或儲存於多個經協調檔案(例如，儲存一或多個模塊、副程式或程式碼之若干部分之檔案)中。一電腦程式可經部署以在一個電腦上或位於一個位點處或跨越多個位點分佈且藉由一資料通信網路互連之多個電腦上執行。

【0083】 一或多個電腦之一系統經組態以執行特定操作或動作意味該系統已將在操作中致使系統執行操作或動作之軟體、韌體、硬體或其一組合安裝於其上。一或多個電腦程式經組態以執行特定操作或動作意味一或多個程式包含在由資料處理設備執行時致使設備執行該等操作或動作之指令。

【0084】 如本說明書中所使用，一「引擎」或「軟體引擎」係指提供不同於輸入之一輸出的一硬體實施或軟體實施之輸入/輸出系統。一引擎可在專用數位電路系統中實施或實施為待由一計算裝置執行之電腦可讀

指令。每一引擎可實施於包含一或多個處理器及電腦可讀媒體之任何適當類型之計算裝置(例如，伺服器、行動電話、平板電腦、筆記型電腦、音樂播放器、電子書閱讀器、膝上型電腦或桌上型電腦、PDA、智慧型電話或其他固定或可攜式裝置)上。另外，引擎中之兩者或兩者以上可實施於同一計算裝置上或不同計算裝置上。

【0085】 本說明書中所闡述之程序及邏輯流程可由執行一或多個電腦程式之一或多個可程式化處理器執行以藉由對輸入資料進行操作並產生輸出來執行功能。該等程序及邏輯流程亦可由特殊用途邏輯電路系統(例如，一FPGA或一ASIC)或由特殊用途邏輯電路系統與一或多個經程式化電腦之一組合執行。

【0086】 適合用於執行一電腦程式之電腦可基於一般或特殊用途微處理器或兩者，或任何其他種類之中央處理單元。一般而言，一中央處理單元將自一唯讀記憶體或一隨機存取記憶體或兩者接收指令及資料。一電腦之基本元件係用於執行指令之一中央處理單元及用於儲存指令及資料之一或多個記憶體裝置。該中央處理單元及該記憶體可由特殊用途邏輯電路補充或併入於特殊用途邏輯電路中。一般而言，一電腦亦將包含用於儲存資料之一或多個大容量儲存裝置(例如，磁碟、磁光碟或光碟)或以操作方式耦合以自該大容量儲存裝置接收資料或向其傳送資料或既接收又傳送資料。然而，一電腦不必具有此類裝置。此外，一電腦可嵌入於另一裝置中，例如，一行動電話、一個人數位助理(PDA)、一行動音訊或視訊播放器、一遊戲控制台、一全球定位系統(GPS)接收器或一可攜式儲存裝置，例如，一通用串列匯流排(USB)隨身碟，僅舉幾個例子。

【0087】 適合於儲存電腦程式指令及資料之電腦可讀媒體包含所有形式之非揮發性記憶體、媒體及記憶體裝置，以實例之方式包含：半導體

記憶體裝置，例如，EPROM、EEPROM及快閃記憶體裝置；磁碟，例如，內部硬碟或可拆卸式磁碟；磁光碟；以及CD-ROM及DVD-ROM碟。

【0088】 為提供與一使用者之互動，本說明書中所闡述之標的物之實施例可實施於一主機裝置上，該主機裝置具有用於向該使用者顯示資訊之一顯示裝置(例如，一CRT (陰極射線管)或LCD (液晶顯示器)監視器)及該使用者可藉以向電腦提供輸入之一鍵盤及指向裝置(例如，一滑鼠、軌跡球或一存在敏感顯示器或其他表面)。亦可使用其他種類之裝置來提供與一使用者之互動；舉例而言，提供給該使用者之回饋可係任何形式之感觀回饋(例如，視覺回饋、聽覺回饋或觸覺回饋)；且來自該使用者之輸入可以任何形式(包含聲音、語音或觸覺輸入)來接收。另外，一電腦可藉由向由一使用者使用之一裝置發送文件及自該裝置接收文件而與該使用者互動；舉例而言，藉由回應於自一使用者之裝置上之一web瀏覽器接收之請求而向該web瀏覽器發送網頁。而且，一電腦可藉由將文字訊息或其他形式之訊息發送至一個人裝置(例如，一智慧型電話，其運行一訊息收發應用程式且繼而接收來自使用者之回應性訊息)與一使用者互動。

【0089】 除上文所闡述之實施例之外，以下實施例亦係創新性的：

【0090】 實施例1係一種周圍計算系統，其包括：一周圍機器學習引擎；一低功率CPU；及一SRAM，其至少在該周圍機器學習引擎及該低功率CPU當中共用；其中該周圍機器學習引擎包括虛擬位址邏輯以自由該周圍機器學習引擎產生之虛擬位址變換為該SRAM內之實體位址。

【0091】 實施例2係如實施例1之周圍計算系統，其中該周圍計算系統經組態以將用於一機器學習模型之參數自與該周圍計算系統分開之一DRAM串流式傳輸至該SRAM中。

【0092】 實施例3係如實施例1至2中任一實施例之周圍計算系統，

其中該周圍計算系統整合至一系統單晶片中，且其中該DRAM與該系統單晶片分開。

【0093】 實施例4係如實施例1至3中任一實施例之周圍計算系統，其中自該DRAM串流式傳輸該機器學習模型之該等參數會重寫先前在該SRAM中之指令或資料。

【0094】 實施例5係如實施例1至4中任一實施例之周圍計算系統，其中該周圍機器學習引擎經組態以藉由自該SRAM讀取該機器學習模型之參數而經由一機器學習模型執行一推理遍次。

【0095】 實施例6係如實施例1至5中任一實施例之周圍計算系統，其中自該SRAM讀取該機器學習模型之該等參數包括由該周圍機器學習引擎執行具有未對應於該SRAM中之位址之虛擬位址的讀取指令。

【0096】 實施例7係如實施例1至6中任一實施例之周圍計算系統，其中該周圍機器學習引擎經組態以將該讀取指令之該等虛擬位址提供至該周圍機器學習引擎以便產生表示該SRAM中之位置之實體位址。

【0097】 實施例8係如實施例1至7中任一實施例之周圍計算系統，其中該周圍機器學習引擎經組態以藉由一編譯程式執行自一輸入程式產生之指令，該編譯程式將可用於該周圍機器學習引擎之局部記憶體之一大小作為一輸入引數，且其中藉由將作為可用之局部記憶體之該SRAM之一大小提供至該編譯程式而產生該等指令。

【0098】 實施例9係如實施例1至8中任一實施例之周圍計算系統，其中該周圍計算系統經組態以執行包括以下各項之操作：接收表示待處理之一或多個感測器信號之接收之一中斷；啟動該一或多個其他處理組件中之一第一處理組件，包含將待由該第一處理組件執行之指令串流式傳輸至該SRAM中；由該第一處理組件使用該SRAM中之該等指令處理該一或多

個感測器信號以判定該周圍機器學習引擎應進一步處理該一或多個感測器信號；啟動該周圍機器學習引擎包含將待由該周圍機器學習引擎使用之參數串流式傳輸至該SRAM中；及由該周圍機器學習引擎使用儲存於該SRAM中之該等參數執行一機器學習模型之一推理遍次。

【0099】 實施例10係如實施例1至9中任一實施例之周圍計算系統，其中將待由該周圍機器學習引擎使用之該等參數串流式傳輸至該SRAM中會覆寫儲存於該SRAM中之由該第一處理組件執行之該等指令。

【0100】 實施例11係如實施例1至10中任一實施例之周圍計算系統，其中與一直接記憶體存取控制器、一或多個其他機器學習引擎或者一或多個其他處理器進一步共用該SRAM。

【0101】 實施例12係如實施例1至11中任一實施例之周圍計算系統，其中該一或多個其他機器學習引擎包括一主要機器學習引擎。

【0102】 實施例13係一種用於虛擬化一周圍計算系統上之記憶體之方法，該系統包括：一周圍機器學習引擎，其包括虛擬位址邏輯；一低功率CPU；及一SRAM，其至少在該周圍機器學習引擎及該低功率CPU當中共用，其中該方法包括：由該周圍機器學習引擎之該虛擬位址邏輯接收由該周圍機器學習引擎產生之虛擬位址；及由該周圍機器學習引擎上之該虛擬位址邏輯將由該周圍機器學習引擎產生之虛擬位址變換為該SRAM內之實體位址。

【0103】 實施例14係如實施例13之用於虛擬化記憶體之方法，其中該周圍計算系統係如實施例1至12中任一實施例。

【0104】 實施例15係編碼有待由一周圍計算系統之一周圍機器學習引擎執行之指令之一或多個電腦可讀儲存媒體，該周圍計算系統進一步包括一低功率CPU及至少在該周圍機器學習引擎及該低功率CPU當中共用之

一SRAM，其中該等指令中之一或多者包括對應於在該周圍機器學習引擎及該低功率CPU當中共用之該SRAM中之實體位址的虛擬位址。

【0105】 實施例16係如實施例15之一或多個電腦可讀儲存媒體，其中該周圍計算系統係如實施例1至12中任一實施例。

【0106】 儘管本說明書含有諸多特定實施細節，但此等細節不應解釋為對任何發明之範疇或對可主張之內容之範疇之限制，而是應解釋為可為特定發明之特定實施例特有之特徵之說明。在單獨實施例之內容脈絡中於本說明書中闡述之特定特徵亦可以組合方式實施於一單個實施例中。相反地，在一單個實施例之內容脈絡中闡述之各種特徵亦可單獨地或以任何適合子組合形式實施於多個實施例中。此外，儘管上文可將特徵闡述為以特定組合形式起作用且甚至最初主張如此，但來自一所主張組合之一或多個特徵在某些情形中可自該組合去除，且該所主張組合可係針對一子組合或一子組合之變化形式。

【0107】 類似地，儘管在圖式中以一特定次序繪示操作，但不應將此理解為需要以所展示之特定次序或以順序次序執行此類操作，或執行所有所圖解說明之操作以達成合意結果。在特定情形下，多任務及並行處理可係有利的。此外，不應將在上文所闡述之實施例中之各種系統模組及組件之分離理解為在所有實施例中需要此分離，且應理解，一般可將所闡述之程式組件及系統一起整合於一單個軟體產品中或封裝至多個軟體產品中。

【0108】 已闡述標的物之特定實施例。在所附申請專利範圍之範疇內存在其他實施例。舉例而言，申請專利範圍中所引用之行動可以一不同次序來執行且仍達成合意結果。作為一項實例，附圖中所繪示之程序未必需要所展示之特定次序或順序次序來達成合意結果。在特定的某些情形

下，多任務及並行處理可係有利的。

【符號說明】

【0109】

100:實例性計算裝置/裝置/計算裝置

102:音訊感測器

104:雷達感測器

106:觸控感測器

108:全球定位系統感測器

109:加速度計

110:周圍計算系統

112:感測器/周邊感測器

114:周邊介面

120:控制子系統

121:功率控制單元

122:功率管理單元

123:時脈控制單元

124:中斷控制器

128:直接記憶體存取控制器

129:定時器

130:處理子系統

132:低功率中央處理單元

134:周圍機器學習引擎

135:虛擬位址邏輯

136:低功率DSP

138:高功率DSP

139:靜態隨機存取記憶體/共用靜態隨機存取記憶體

140:主要中央處理單元叢集

150:主要機器學習引擎

200:系統單晶片

210: DRAM

220:記憶體控制器

230:系統級快取記憶體

240:系統單晶片組構

250:計算裝置組件

252:通信路徑

254:通信路徑

310:操作

320:操作

330:操作

【發明申請專利範圍】

【請求項1】

一種用於虛擬化外部記憶體之裝置，該裝置包括：

一主要記憶體，其由多個用戶端(client)裝置共用，其中該多個用戶端裝置包含一周圍(ambient)計算裝置；

其中該周圍計算裝置包括多個周圍處理裝置及由該多個周圍處理裝置所共用(shared)之一共用局部記憶體，其中該多個周圍處理裝置包含一周圍機器學習(ML)引擎；

其中該周圍計算裝置包含虛擬位址邏輯(virtual address logic)，其經組態以將藉由該周圍ML引擎所使用之虛擬位址變換(translate)為該共用局部記憶體之實體位址。

【請求項2】

如請求項1之裝置，其中在共用該主要記憶體之其他用戶端裝置自一低功率狀態啟動之前，該周圍計算裝置經組態以處理感測器信號。

【請求項3】

如請求項1之裝置，其中該多個用戶端裝置包含一主要ML引擎，其在該主要記憶體中產生實體位址。

【請求項4】

如請求項1之裝置，其中在接收一中斷之後，該裝置經組態以將模型參數自該主要記憶體串流式傳輸(stream)至該共用局部記憶體中。

【請求項5】

如請求項4之裝置，其中將該等模型參數串流式傳輸至該共用局部記憶體中重寫(overwrite)藉由一或多個其他周圍處理裝置所使用之該共用局

部記憶體中之空間。

【請求項6】

如請求項4之裝置，其中將該等模型參數串流式傳輸至該共用局部記憶體中包括將該等模型參數串流式傳輸至該共用局部記憶體中之可用空間中。

【請求項7】

如請求項4之裝置，其中該中斷表示待處理之一或多個感測器信號之接收。

【請求項8】

如請求項4之裝置，其中該周圍ML引擎經組態以藉由使用該虛擬位址邏輯以存取經串流式傳輸至該共用局部記憶體中之該等模型參數以執行一機器學習模型之一推理遍次(inference pass)。

【請求項9】

如請求項1之裝置，其中該共用局部記憶體包括多個記憶體組(bank)，該多個記憶體組經組態以當進入一低功率狀態時，個別地斷開電源(powered down)。

【請求項10】

如請求項1之裝置，其中該多個周圍處理裝置包括一直接記憶體存取控制器、一或多個其他ML引擎或者一或多個處理器中之至少一者。

【請求項11】

如請求項1之裝置，其中該周圍ML引擎包括一單一ML計算塊(compute tile)。

【請求項12】

一種用於虛擬化外部記憶體之系統，該系統包括：

多個用戶端裝置，其包含一周圍計算裝置；及

一主要記憶體，其由該多個用戶端裝置共用；

其中該周圍計算裝置包括多個周圍處理裝置及由該多個周圍處理裝置所共用之一共用局部記憶體，其中該多個周圍處理裝置包含一周圍機器學習(ML)引擎；

其中該周圍計算裝置包含虛擬位址邏輯，其經組態以將藉由該周圍ML引擎所使用之虛擬位址變換為該共用局部記憶體之實體位址。

【請求項13】

如請求項12之系統，其中在共用該主要記憶體之其他用戶端裝置自一低功率狀態啟動之前，該周圍計算裝置經組態以處理感測器信號。

【請求項14】

如請求項12之系統，其中該多個用戶端裝置包含一主要ML引擎，其在該主要記憶體中產生實體位址。

【請求項15】

如請求項12之系統，其中在接收一中斷之後，該裝置經組態以將模型參數自該主要記憶體串流式傳輸至該共用局部記憶體中。

【請求項16】

如請求項15之系統，其中將該等模型參數串流式傳輸至該共用局部記憶體中重寫藉由一或多個其他周圍處理裝置所使用之該共用局部記憶體中之空間。

【請求項17】

如請求項15之系統，其中將該等模型參數串流式傳輸至該共用局部

記憶體中包括將該等模型參數串流式傳輸至該共用局部記憶體中之可用空間中。

【請求項18】

如請求項15之系統，其中該中斷表示待處理之一或多個感測器信號之接收。

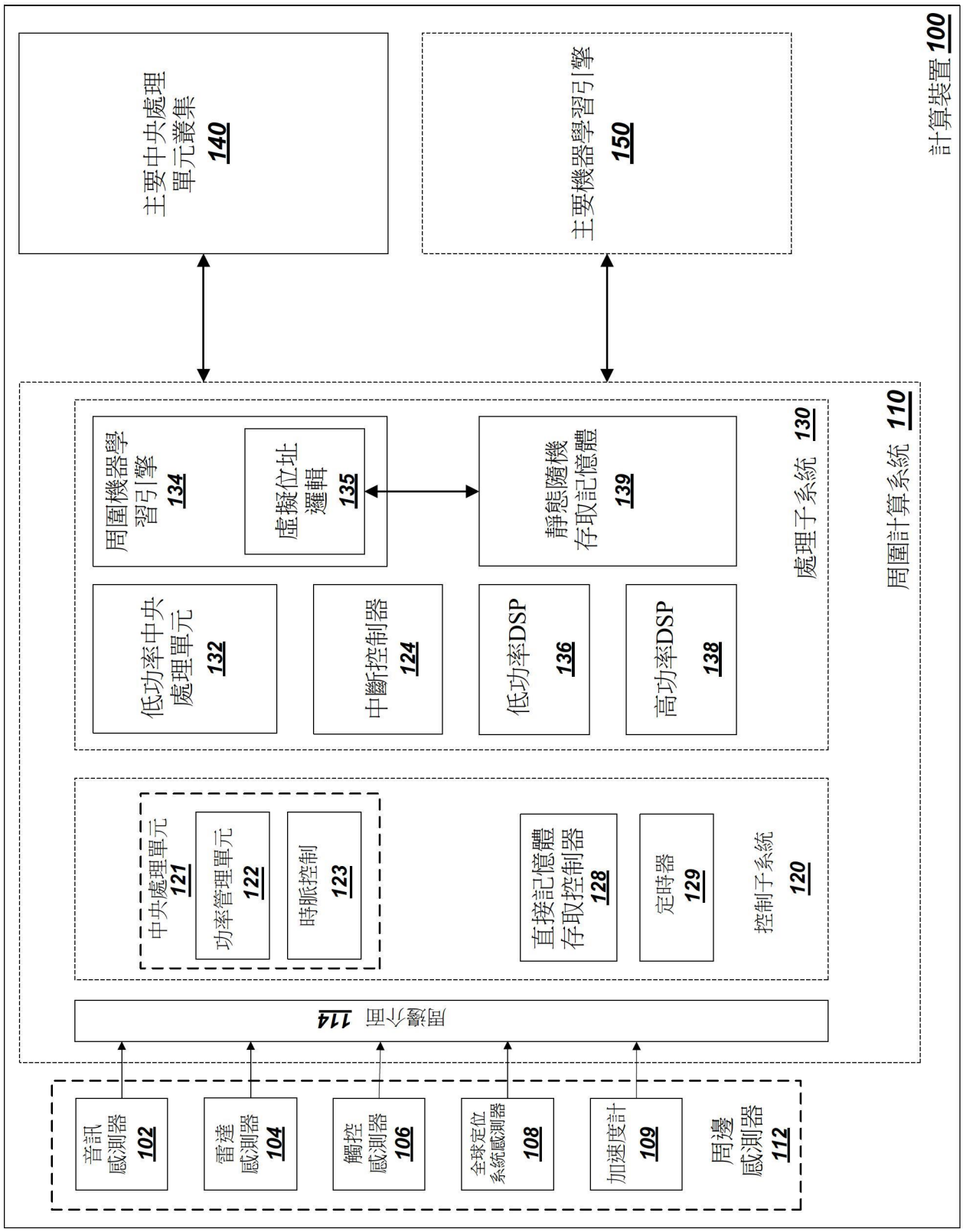
【請求項19】

如請求項15之系統，其中該周圍ML引擎經組態以藉由使用該虛擬位址邏輯以存取經串流式傳輸至該共用局部記憶體中之該等模型參數以執行一機器學習模型之一推理遍次。

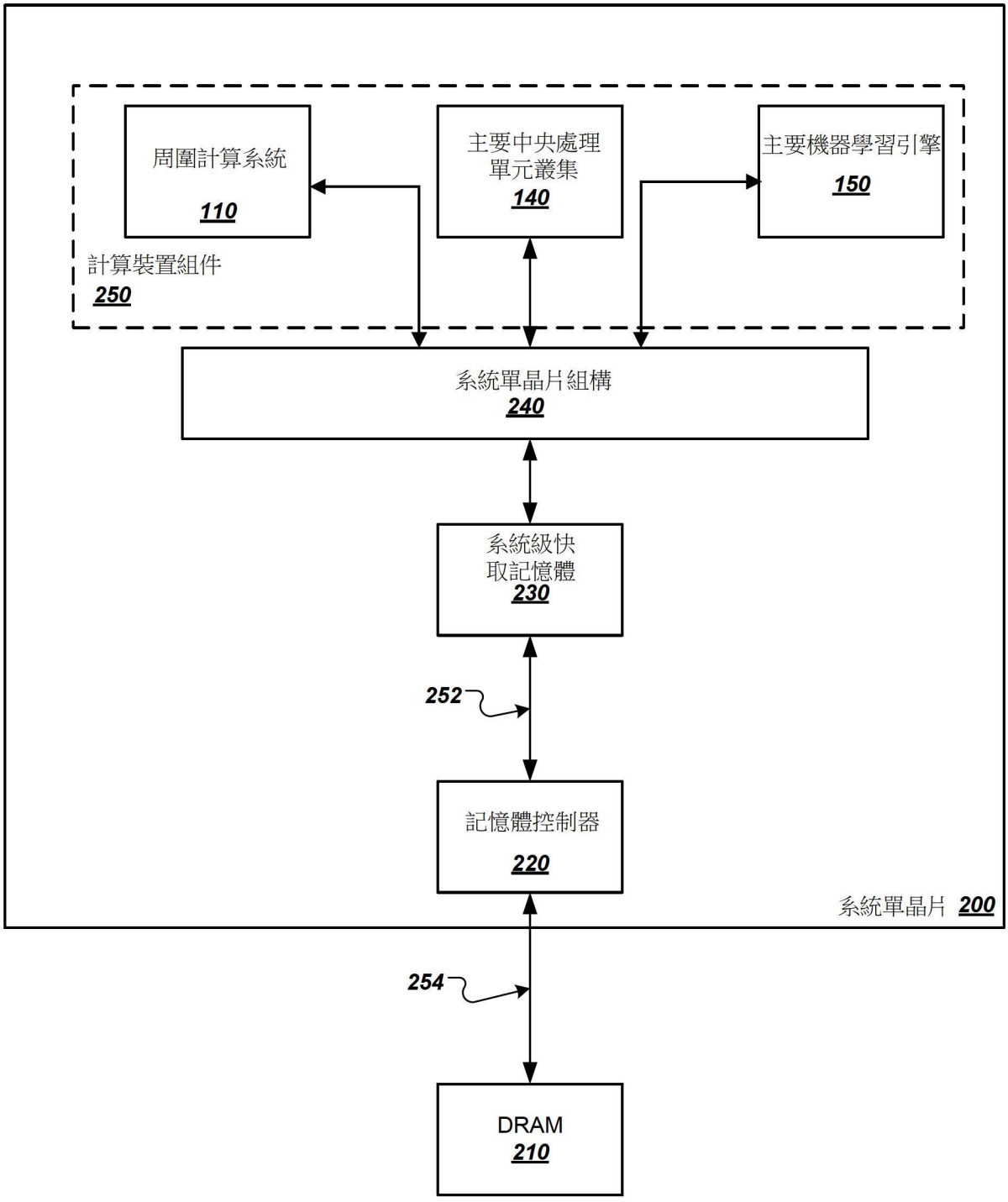
【請求項20】

如請求項12之系統，其中該共用局部記憶體包括多個記憶體組，該多個記憶體組經組態以當進入一低功率狀態時，個別地斷開電源。

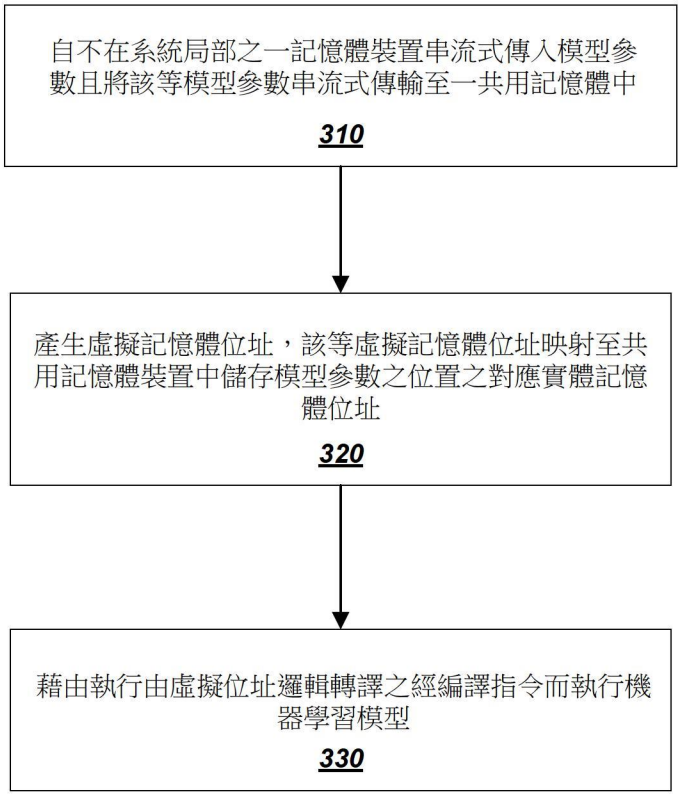
【發明圖式】


計算裝置 100
主要中央處理單元叢集 140
主要機器學習引擎 150
周圍計算系統 110
處理子系統 130
周圍機器學習引擎 134
虛擬位址邏輯 135
靜態隨機存取記憶體 139
低功率中央處理單元 132
中斷控制器 124
低功率DSP 136
高功率DSP 138
控制子系統 120
中央處理單元 121
功率管理單元 122
時脈控制 123
直接記憶體存取控制器 128
定時器 129
周邊介面 114
周邊感測器 112
音訊感測器 102
雷達感測器 104
觸控感測器 106
全球定位系統感測器 108
加速度計 109

【圖1】


周圍計算系統
110
主要中央處理
單元叢集
140
主要機器學習引擎
150
計算裝置組件
250
系統單晶片組構
240
系統級快
取記憶體
230
252
記憶體控制器
220
系統單晶片 200
254
DRAM
210

【圖2】


自不在系統局部之一記憶體裝置串流式傳入模型參數且將該等模型參數串流式傳輸至一共用記憶體中
310
產生虛擬記憶體位址，該等虛擬記憶體位址映射至共用記憶體裝置中儲存模型參數之位置之對應實體記憶體位址
320
藉由執行由虛擬位址邏輯轉譯之經編譯指令而執行機器學習模型
330

【圖3】