



(21) 申請案號：097143884

(22) 申請日：中華民國 97 (2008) 年 11 月 13 日

(51) Int. Cl. : **G06F15/80 (2006.01)**

(30) 優先權：2007/12/07 美國 11/952,828

(71) 申請人：微軟公司 (美國) MICROSOFT CORPORATION (US)
美國

(72) 發明人：范德霍文阿里 VAN DER HOEVEN, ARIE (US)；沃克愛爾斯沃斯 D WALKER, ELLSWORTH D. (US)；福滋福瑞思 C FOLTZ, FORREST C. (US)；鄧忠 DENG, ZHONG (US)

(74) 代理人：蔡坤財；李世章

(56) 參考文獻：

TW	200709100A	US	5872963
US	7111188B2	US	2004/0019891A1
US	2005/0149603A1	US	2007/0039002A1

審查人員：鄭書季

申請專利範圍項數：19 項 圖式數：2 共 0 頁

(54) 名稱

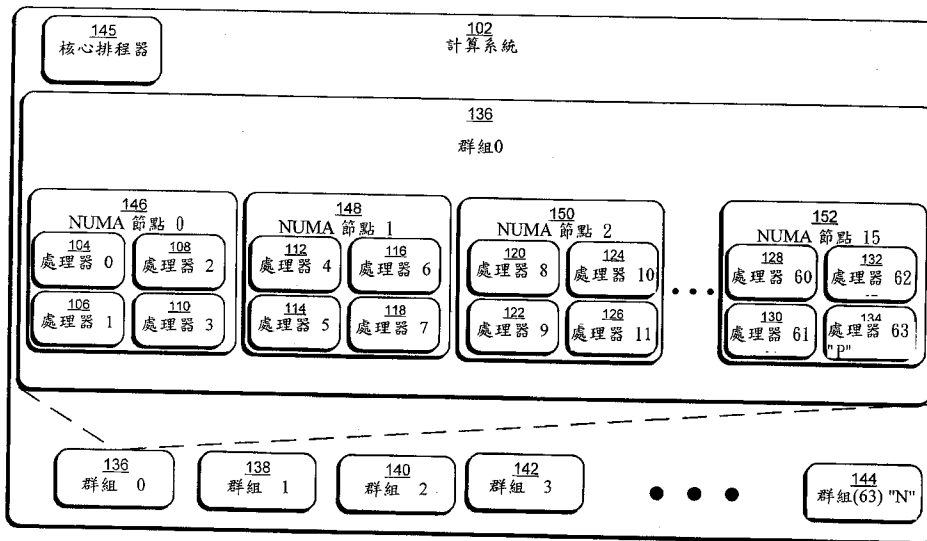
用於核心處理器分組的方法、系統及電腦可讀取媒體

METHOD, SYSTEM, AND COMPUTER-READABLE MEDIA FOR KERNEL PROCESSOR GROUPING

(57) 摘要

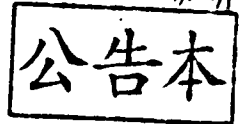
揭示將個別處理器分組成各指派實體的技術。靜態地分組各處理器可允許在一群組基礎上指派各執行緒。在此方式中，排程各執行緒以供處理的責任可被減小，同時該指派實體內的處理器可基於該群組內之該等個別處理器之實體區域性(locality)而被選擇。該分組允許一系統來調整規模(scale)以符合各種應用程式的處理需求。

Techniques for grouping individual processors into assignment entities are discussed. Statically grouping processors may permit threads to be assigned on a group basis. In this manner, the burden of scheduling threads for processing may be minimized, while the processor within the assignment entity may be selected based on the physical locality of the individual processors within the group. The groupings may permit a system to scale to meet the processing demands of various applications.



102 . . . 計算系統
 145 . . . 核心排程器
 136 . . . 群組 0
 104-132 . . . 處理器
 0-62
 134 . . . 處理器
 63''P''
 146-152 . . . NUMA
 節點 0-15
 136-144 . . . 群組
 0-63

第1圖



發明專利說明書

(本說明書格式、順序、請勿任意更動，※記號部分請勿填寫；惟已有申請案號者請填寫)

※ 申請案號：97143884

※ 申請日期：2008 年 11 月 13 日

※IPC 分類：G06F 15/80 (2006.01)

一、發明名稱：(中文/英文)

用於核心處理器分組的方法、系統及電腦可讀取媒體/METHOD,
SYSTEM, AND COMPUTER-READABLE MEDIA FOR KERNEL
PROCESSOR GROUPING

二、中文發明摘要：

揭示將個別處理器分組成各指派實體的技術。靜態地分組各處理器可允許在一群組基礎上指派各執行緒。在此方式中，排程各執行緒以供處理的責任可被減小，同時該指派實體內的處理器可基於該群組內之該等個別處理器之實體區域性 (locality) 而被選擇。該分組允許一系統來調整規模 (scale) 以符合各種應用程式的處理需求。

三、英文發明摘要：

Techniques for grouping individual processors into assignment entities are discussed. Statically grouping processors may permit threads to be assigned on a group basis. In this manner, the burden of scheduling threads for processing may be minimized, while the processor within the assignment entity may be selected based on the physical locality of the individual processors within the group. The groupings may permit a system to scale to meet the processing demands of various applications.

四、指定代表圖：

(一)本案指定代表圖為：第(1)圖。

(二)本代表圖之元件符號簡單說明：

102 計算系統

145 核心排程器

136 群組 0

104-132 處理器 0-62

134 處理器 63”P”

146-152 NUMA 節點 0-15

136-144 群組 0-63

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

無

六、發明說明：

【發明所屬之技術領域】

本發明係關於核心處理器分組。

【先前技術】

隨著系統中處理器的增加，當處理經設計來在具有較少處理器之一系統上被處理的應用程式時，該等處理器的整體生產力不符合該系統期望的處理能力。例如，當針對處理的個別執行緒被分配至各種處理器時，將會發生瓶頸。在其他範例中，特定應用程式限制可有效處理該應用程式之任務之處理器的數目。例如，特定應用程式不良好合適於由多個處理器處理，基於此而設計該應用程式。例如，當一應用程式如期望地操縱在包含兩個處理器的一桌上系統時，具有六十四個或更多個處理器的一企業伺服器可能遭遇到相同應用程式的問題。

【發明內容】

揭示將個別處理器分組成各指派實體的技術。靜態地分組各處理器可允許在一群組基礎上指派各執行緒。在此方式中，排程各執行緒以供處理的責任可被減小，同時該指派實體內的處理器可基於該群組內之該等個別處理器之實體區域性（locality）而被選擇。該分組允許一系統來調整規模（scale）以符合各種應用程式的處理需求。

此發明內容係提供來引介在一簡化形式中的概念選項，其將於如下實施方式中更進一步說明。此發明內容並不意圖用來識別所請標的之關鍵特徵或必要特徵，亦不意圖來被用作為決定所請標的之範疇。

【實施方式】

概述

描述可提供核心階層處理器分組的各技術。例如，基於各個別處理器之區域性，可靜態地將個別處理器組態成核心階層分組，以致使用於處理之執行緒或分離的應用程式可在一每一群組基礎上被排程且處理。分組可允許作業系統來在一群組基礎上進行互動，而非與個別處理器互動。因此，相較於在每一處理器基礎上分配處理任務的作業系統，上述作業系統可較為簡化。該個別執行緒可被指派至一核心群組以供處理。靜態地分組處理器且在一每一分組群組上指派處理任務，其可減小與在包含大量處理器之系統中排程個別處理器的責任。例如，具有一百二十八個處理器之一企業伺服器可經組態來處理被設計來在一具兩個處理器之桌上類型系統中有效操作的多個應用程式。

對可遭遇同步化或正確性問題的應用程式來說，如果經由被期望來處理該應用程式之多個處理器來實現處理時，與該應用程式相關的執行緒可在一單一核心分組中被處理，以便隔離處理（例如，發生猶如在核心群組中之該等處理器為在該系統中之該唯一處理器）。在此方式中，一第一核心群組可服務一第一應用程式，同時，如果適當的話，其他應用程式可由其他核心群組來處理。該靜態處理器分組可促使有效處理經設計來在有限數量處理器上進行處理的應用程式，同時支援運行在該計算系統上之該等應用程式的整體可擴充性（scalability）。

範例式環境

第 1 圖描述一允許核心分組的範例實施中的環境 100。例如，具有六十四或更多個處理器（104-134 將被

參照) 的一計算系統 102 可經組態，以致使該等處理器被組態成各群組 (群組「0-N」將被參照，個別地指稱為 136-144) 於該核心階層上。作業在該計算系統上之一作業系統可在該核心階層上經組態以造成執行緒或分離的任務來在該計算系統中之各處理器之特定群組上被處理。核心階層處理器分組可減輕必須去負責包含在該系統中之處理器數目同時考量該等個別處理器的實體區域性之各應用程式或者其他階層。也就是說，該些經由負責該核心階層上的多重處理器者，相較於該些基於組態一應用程式者，關於在多個處理器上進行處理的問題可被減小。例如，軟體應用程式模組可被隔離或另分割以作出該可用處理資源的使用。因此，非立即可調整規模之各應用程式可經處理而，相較於一可預料處理器使用，不需過度地消耗計算系統資源。例如，當在大量處理器上處理時，一應用程式可消耗的多個處理能力，其係相較於當在針對設計該應用程式之處理器數目上運行時的相同應用程式。在一每一群組基礎上執行各操作可減小當分配執行緒以供處理時在該作業系統上的排程責任。例如，經由分配個別執行緒至一核心群組以供處理，在一核心排程器 145 上的責任可相較於分配各任務至個別處理器而相對減小。例如，減小該計算系統中的核心群組的數目可允許該核心排程器 145 來實行一相對簡單且潛在性快速的演算法。

當討論實質具體實施例時，虛擬性實施也應可被瞭解。例如，應用程式可運作在一虛擬環境中、或在一組合環境中、等等。例如，一應用程式可執行在一也實質上被分割的計算系統上的一虛擬環境中。

雖然，一 64 (六十四) 處理器計算系統被參照，此中所討論的該技術及原理也可被調整規模，例如針對「P」處理器 (例如在一特定群組中) 以及「N」核心群組 (包

含「P」個別處理器)，其可如所期望地基於計算系統資源，例如一百二十八個處理器系統、硬體/軟體、等等。例如，該六十四個或更多個處理器可經組態成「N」核心階層分組以及具有「P」處理器的個別核心分組。在實施中，一核心群組包含一單一處理器、兩處理器、等等。在另一範例中，當一六十四位元 bitmark (位元遮罩) 可有效定址該分組時，一六十四處理器系統可被組態為一單一群組。在更進一步的範例中，具有六十四或更多個處理器的一系統可被組態成每一群組大概六十四處理器。

核心群組及在個別核心群組內之個別處理器的精確數目可變化。例如，處理器可被熱交換 (hot-swapped) 成所期望的群組、等等。例如，當一計算系統包含六十四或更多個實體處理器時，特定群組可專屬於特定任務，或者處理器可為接續群組指派而保留，等等。排程任務之責任可經由減小核心群組之數目而被限制。例如，一群組內之個別處理器的數目可基於設計一應用程式來在其上被處理的處理器之數目而被選擇。在此方式中，一計算系統可支援不自我提供至由大量處理器的處理的應用程式。例如，如果一應用程式 (預期來被常規地處理) 將有效地利用兩處理器，一核心群組可包含兩處理器。也就是說，在核心群組中之該個別處理器的數目可基於應用程式參數或者針對預期來被處理之應用程式而被指派。例如，如果一企業伺服器預料來執行經設計來在四個處理器上處理的至少一應用程式，四個處理應用程式可靜態地指派至一核心群組，以致使可被有效利用的額外處理器不包含在該核心群組中。

其他考量可包含一起分組一足夠的處理器數目來有效處理一執行緒。除了排程應用程式執行緒至專屬群組外，該核心排程器 145 可當核心群組變成可獲用於處理

時，在一循環 (round-robin) 基礎上，指派各任務。

核心群組內之個別處理器可基於與該核心群組內之其他處理器相關之該處理器的區域性來被選擇。在實施中，非一致性記憶體存取 (non-uniform memory access, NUMA) 內之個別處理器可被包含在核心群組中。例如，指派至一本地記憶體節點的一或更多個別處理器可被包含在核心群組中。因此，當處理指派至包含該 NUMA 節點中之該等處理器的該核心群組的一執行緒時，該 NUMA 節點內之該等處理器可有效存取本地記憶體資源。依次地 (In-turn)，包含在該 NUMA 節點內 (如 NUMA 節點 0-15，個別標示為 146-152) 之該等處理器可被指派至特定核心分組。包含一核心分組內之區域性相鄰處理器 (locality adjacent processors)，無論是否在相同 NUMA 節點中，可增進整體處理同時該系統之各種部分可由不同應用程式所使用。用於決定分組之其他因素可關聯於區域性而被使用或者可被實施，例如處理核心組態或其他期望的組態。例如，一核心分組組態可基於一處理核心之組態以及該核心的插座組態。

核心階層處理器分組可避免相較於針對設計一應用程式之一較少處理器系統而用於具有大量處理器數目之計算系統的不穩定的應用程式效能、正確性問題、同步化問題、等等。例如，具有大量處理器數目 (例如一百二十八個) 之一計算系統可能遭遇上述問題，但是運作該相同程式之一較小資源系統將不會。根據此中技術之分組該處理器可造成該應用程式及/或該系統有效地減小不穩定行為的可能性。

在實施中，一核心內之處理器的數目允許一經選定大小的一共通位元遮罩來被使用。例如，一六十四位元遮罩可在一有效方式中被管理，同時符合該經分組的處理器組態。其他範例可實施一 32 位元位元遮罩組態。

經由使用核心群組，指派用於處理之各執行緒的作業系統可有效使用計算系統資源，藉以避免發生在具有大量處理器數目之伺服器中的可能性問題。在一大規模處理器系統中，一核心排程器 145 可指派用於處理之個別應用程式執行緒至特定核心群組，以致使多重應用程式（相較於包含在該系統中之其他者，其合適於一較少量的處理器）可在一般同時性的方式中經處理，藉以相較於該等處理器在一個別基礎上經管理而較有效作出使用該系統的處理器。

如果一特定應用程式具有同步化，正確性或其他多重處理器問題（如果執行在一具有多重處理器的計算系統），該核心排程器 145 可排程該應用程式的執行緒至一單一群組來避免或減小該些可能的問題。例如，如果一應用程式不具立即可擴充性，該核心排程器 145 可導引該應用程式之處理任務至一單一群組。例如，如果一具有六十四或更多個處理器的電腦系統係處理為多重處理器性的一任務，該核心排程器 145 可導引該應用程式執行緒至群組 0（其隔離該應用程式）。在此方式中，該等執行緒可被處理猶如在該群組中之該等處理器為該處理系統的處理資源。隔離的階層可從一實體或虛擬分割類型隔離至如所期望的較少形式之隔離來作變化。

相對地，如果一應用程式經組態用於多重處理器處理，用於處理之各執行緒可被個別地排程以供在群組 1、群組 2、及群組 3 之間（其個別地可包含相近於群組 0 之組態的多重處理器）進行處理，藉以有利於該計算系統的處理器資源。

在實施中，各應用程式及驅動程式可被帶給至該整體系統的可見度。例如，一驅動程式可瞭解該核心階層結構，藉以該驅動程式可支援可存取該系統的一部件。在此方式中，該計算系統可獲得群組處理效益，同時應用

程式以及驅動程式可瞭解該系統處理器分組。

一般來說，此中所描述之任何功能可使用軟體、韌體、硬體（例如固定邏輯電路）、手動處理、或上述實施之組合來實作。此中所使用之該項「模組」、「功能性」、及「邏輯」一般可表示為軟體、韌體、硬體、或上述實施之組合。例如一軟體實施，該模組功能性或邏輯表示當執行在一（多）處理器上時所執行特定任務的程式碼。該程式碼能夠經儲存在一或更多電腦可讀記憶體裝置中，例如有形記憶體、等等。

該下列的討論描述利用該先前描述的系統及裝置而實行的轉換技術。該等過程之各者的態樣可按軟體、韌體、硬體、或上述實施之組合而被實作。該等過程經顯示為指示由一或更多裝置所執行之操作的一系列方塊，且其不需被限制於顯示用於執行由該個別方塊之各操作的順序。

範例式過程

該下列的討論描述可利用該先前描述的系統及裝置而實作的一方法。該等過程該等過程之各者的態樣可按軟體、韌體、硬體、或上述實施之組合而被實作。該等過程經顯示為指示由一或更多裝置所執行之操作的一系列方塊，且其不需被限制於顯示用於執行由該個別方塊之各操作的順序。也可體會到各種其他範例。

第 2 圖描述用於靜態分組處理器之範例式過程。例如，一計算系統的作業系統核心階層可經組態至群組個別處理器，因此，可處理敏感於大規模處理器環境的應用程式。

此中所討論之技術可允許在一群組基礎上且在一隔離方式中處理由該作業系統所指派的各執行緒。該些技術可當處理被考量在一群組基礎上而非在一個別處理器基礎上分析出任務時減小該作業系統的複雜度。

針對大等級處理所設計的應用程式，該個別群組可隔離在一指派實體中之該個別任務於該計算系統內之其他指派實體。

該個別處理器可被群組化成一指派實體用於處理執行緒（202）。例如，該核心排程器可指派一特定應用程式至在啟動時所靜態群組化的各處理器。例如，來自一第一應用程式的各執行緒可經排程至一第一指派實體，同時其他應用程式任務可經指派至一第二指派實體、等等。例如，一第一應用程式可經指派至兩個處理器的一第一靜態核心分組，其可實質上本地於包含在該群組中之其他處理器，同時一第二應用程式經指派至具有四個處理器的一第二靜態核心分組。該前述可允許該第一及第二應用程式來被處理地更有效率（作出有效使用該等處理器）猶如該作業系統在一個別基礎上與該等處理器進行互動。

在實施中，包含在一指派實體中的處理器數目可同量於組態一應用程式的處理器數目。在此方式中，處理資源不被專屬於一特定應用程式（其不能夠做出有效使用指派至該群組之處理器之數目）。其他指派實體可相似地經組態，以致使該個別群組可被指派個別來自不同應用程式之各任務。

包含在一指派實體中的處理器數目可基於關於該等處理器而使用的該位元遮罩而被決定。因此，例如，一群組內之該處理器數目可為六十四個或以下，藉以作出一六十四位元位元遮罩的使用。在此方式中，該系統可使用較低位元位元遮罩組態，其可為一簡單組態，同時符合具有超過該位元遮罩所能有效定址的處理器之一系統。例如，當特定處理器被抑制熱交換以及等等時，指派至處理群組之處理器數目可小於該位元遮罩的處理器數目。

使用一群組化處理器組態（在該核心階層上）可減小執行在一較高階層之應用程式上之一多重處理器環境的影響。因此，當該核心階層經組態用於控制多重處理器作為一實體，例如在一真實環境中、一虛擬環境中、或上述之組合，一較低位元位元遮罩組態可經使用而無須重新組態該位元遮罩（針對超過組態該位元遮罩之額外處理器）。

在實施中，一作業系統/計算系統內之指派實體的數目可為了減小與分散應用程式任務（例如，可能性瓶頸）相關之排程責任而被減小。相較於個別定址該等處理器之一系統，此可允許使用較簡化演算法的核心階層排程。指派實體的精確數目以及該指派實體內之個別處理器的數目可基於許多因素，其包含該期望的應用程式處理需求、在該應用程式遭遇處理問題之前，與一應用程式所實施之處理器數目、減小調整規模之瓶頸、等等。

個別處理器可被包含在一特定指派實體中來最大化該指派實體內之該等處理器的區域性。例如，一指派實體內之該等處理器可實質上鄰近於該指派實體內之該其他個別處理器。

一非一致性記憶體存取點（NUMA）節點內的個別處理器可經分組成一指派實體。因此，在一 NUMA 節點中所指派的個別處理器可經分組成一特定核心階層處理器群組（204）。在先前方式中，包含在該指派實體中以及該 NUMA 中的個別處理器可快速存取該本地記憶體（相較於其中該前者所不應用的個別處理器）。當 NUMA 節點不被繫（tied to）至特定核心群組時，在實施中，各節點可被映射至核心群組（206），藉以親近關聯（closely affinitize）於特定任務。例如，一應用程式介面可經組態以將個別實體處理器與該核心分組相繫。親近地映射實體資源與該核心階層指派實體可允許親近地指派相關

任務至特定核心分組以供處理。

針對適於調整規模之應用程式，該核心排程器可指派如所期望的該等執行緒（208）。對針對一限制的處理器數目的應用程式來說，該等執行緒可導向至一單一核心群組。例如，一應用程式可具有導向至群組 0 之處理任務，同時其他應用程式可被導向至群組 1。

應用程式及驅動程式可被帶給至該系統的分組的可見度。例如，一驅動程式可瞭解該核心階層分組結構，藉以該驅動程式可支援可存取該整體系統的一部件（210）。在此方式中，符合此中所討論之該等技術而操作之一計算系統可獲得群組處理效益，同時應用程式以及驅動程式可瞭解所期望之該系統處理器分組。

結論

雖然本發明已經在特定於結構特徵及/或方法邏輯動作之語言中描述，應可瞭解到定義在隨附申請專利範圍中之本發明不須限制於所描述特定特徵或動作。然而，該等特定特徵及動作係揭示作為實施所請發明之範例形式。

【圖式簡單說明】

上述實施方式係參照隨附圖式而加以描述。

第 1 圖描述可使用核心階層分組處理之範例式實施中的一環境；及

第 2 圖描述使用處理器指派實體靜態分組之範例式實施中之一過程的一流程圖。

【主要元件符號說明】

102 計算系統

- 145 核心排程器
 - 136 群組 0
 - 104-132 處理器 0-62
 - 134 處理器 63”P”
 - 146-152 NUMA 節點 0-15
 - 136-144 群組 0-63
 - 202 分組處理器成核心階層指派實體
 - 204 指派區域性相關處理器至 NUMA 節點
 - 206 映射核心階層分組至系統實體參數
 - 208 調整分組規模以適於系統資源
 - 210 加入適於調整規模的處理之裝置的驅動程式模型
- 擴充

七、申請專利範圍：

1. 一種用於核心處理器分組的方法，該方法包含：

在一核心階層（kernel level）上靜態地將個別處理器分組成一或更多個指派實體，該等個別處理器係基於該等個別處理器相對於該群組內之其他個別處理器之實體區域性（locality）而被分組；及

調整該一或更多個指派實體之規模來管理個別執行緒，以致使該等個別執行緒以每一指派實體為基礎被指派。
2. 如申請專利範圍第 1 項所述之方法，其中

該靜態地分組步驟包括靜態地將該等個別處理器之至少一些分組成該一或更多個指派實體之一者，該等個別處理器之該至少一些的數量對應於一應用程式被設計執行於其上之處理器的數量，以及

該調整規模步驟包括調整該一或更多個指派實體之該一者的規模以管理該應用程式之執行緒。
3. 如申請專利範圍第 1 項所述之方法，其中該一或更多個指派實體之至少一者係經組態以將與一應用程式相關之執行緒隔離在該一或更多個指派實體之該至少一者內。
4. 如申請專利範圍第 1 項所述之方法，其中每次指派一單一執行緒至一單一指派實體。
5. 如申請專利範圍第 1 項所述之方法，其中該一或更多個指派實體之一者內之個別處理器係經組態成與該一或更多個指派實體之該一者內之其他個別處理器成為

一非一致性記憶體存取 (non-uniform memory access (NUMA)) 節點。

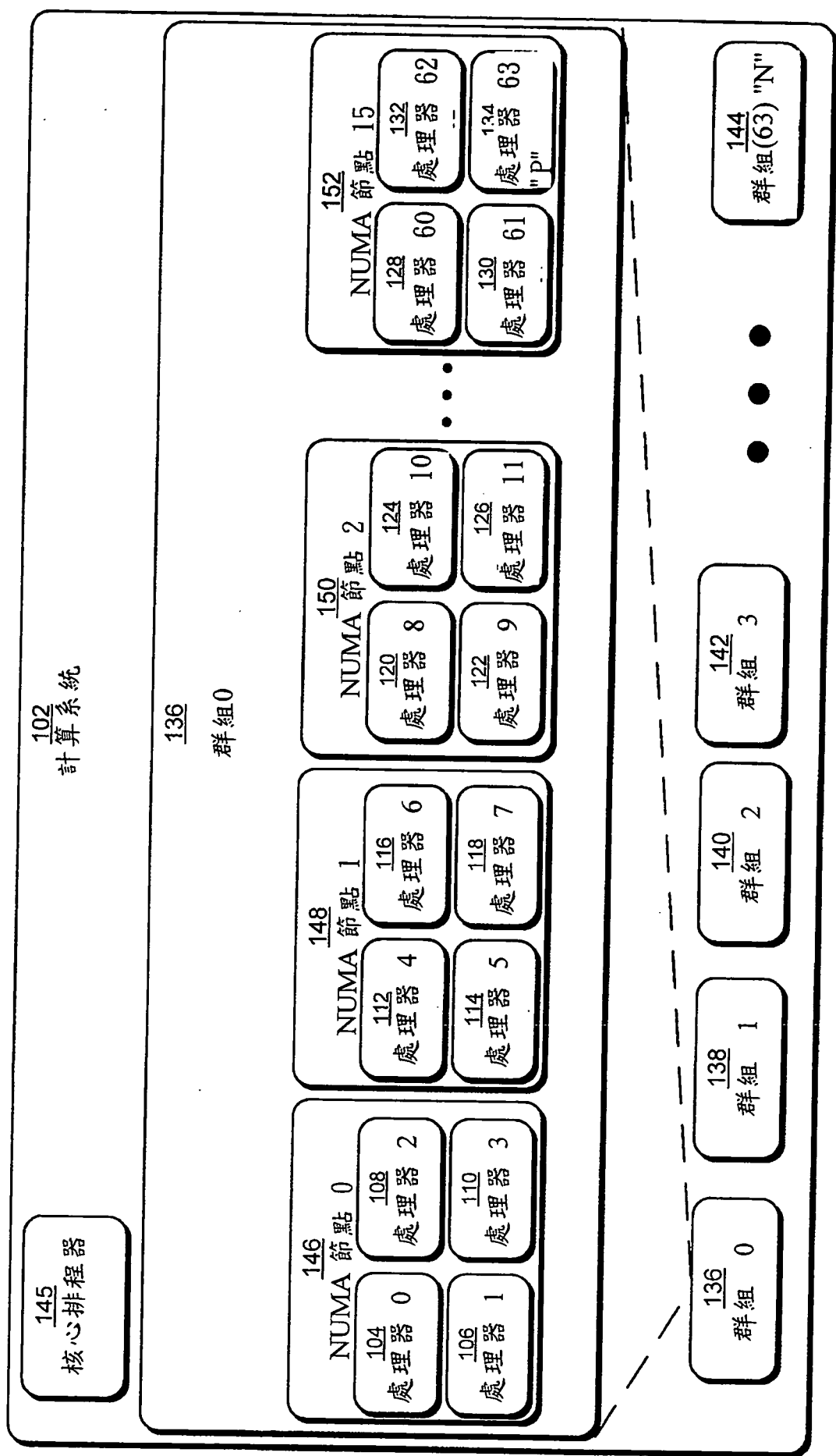
6. 如申請專利範圍第 1 項所述之方法，其中該等個別處理器在啟動時經分組。
7. 如申請專利範圍第 1 項所述之方法，該方法更包含：
在啟動時部分地填入 (partially populating) 一非一致性記憶體存取節點，藉以允許熱加入 (hot-adding) 處理器。
8. 如申請專利範圍第 1 項所述之方法，其中一單一應用程式之各執行緒在一單一指派實體中被處理。
9. 一種包含電腦可執行指令的電腦可讀媒體，當執行該等電腦可執行指令時，該等電腦可執行指令造成一計算系統進行下列步驟：
指派區域性 (locality) 相關之個別處理器至非一致性記憶體存取節點；及
在啟動時靜態地將該等區域性相關之個別處理器分組成一核心群組，該核心群組經組態以每次處理一個別執行緒，該等區域性相關之個別處理器係在一核心階層上分組。
10. 如申請專利範圍第 9 項所述之電腦可讀媒體，其中該等電腦可執行指令更造成該電腦系統進行下列步驟：
以每一核心群組為基礎來排程該個別執行緒。
11. 如申請專利範圍第 9 項所述之電腦可讀媒體，其中應用程式執行緒都被排程至一單一核心群組以供處理。

12. 如申請專利範圍第 9 項所述之電腦可讀媒體，其中該等電腦可執行指令更造成該電腦系統進行下列步驟：
加入經組態用於可調整規模的處理之硬體的驅動程式模型擴充。
13. 如申請專利範圍第 9 項所述之電腦可讀媒體，其中該等電腦可執行指令更造成該電腦系統進行下列步驟：
調整額外的核心群組之規模以供處理來自支援各執行緒分配在不同核心群組中之應用程式的執行緒。
14. 如申請專利範圍第 9 項所述之電腦可讀媒體，其中該核心群組中之個別處理器係與該核心群組外的處理器隔離。
15. 如申請專利範圍第 9 項所述之電腦可讀媒體，其中該核心群組包含大概六十四個個別處理器。
16. 一種用於核心處理器分組的系統，該系統包含：
複數個處理器，該等複數個處理器係在一核心階層上經靜態地組態於一核心群組中，以致使用於處理的個別執行緒以一核心群組為基礎被指派，該等複數個處理器彼此實體相近。
17. 如申請專利範圍第 16 項所述之系統，該系統更包含一核心排程器，該核心排程器係經組態以一核心群組為基礎來分配該等個別執行緒。
18. 如申請專利範圍第 17 項所述之系統，其中該核心排程器指派一特定應用程式之所有執行緒至一單一核心群

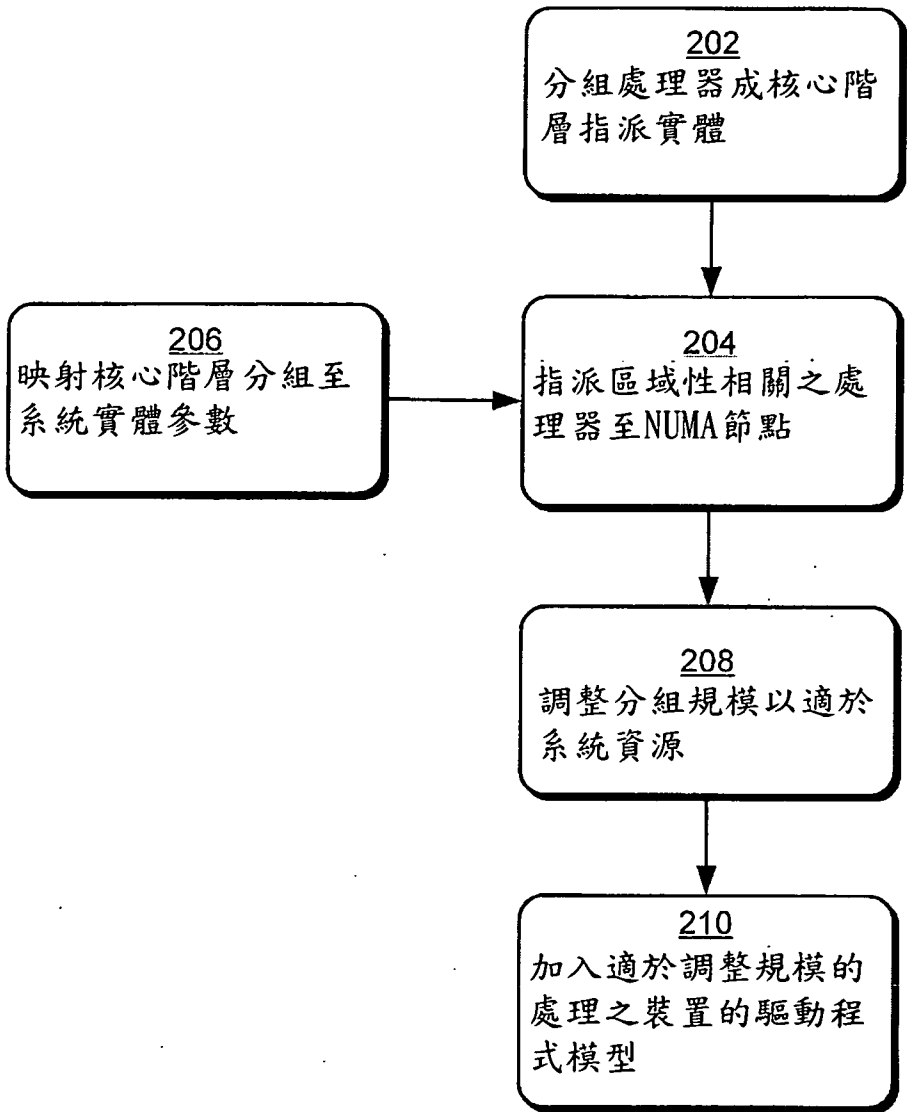
組。

19. 如申請專利範圍第 16 項所述之系統，其中該等複數處理器在啟動該系統時經指派至一核心群組。

+



第1圖



第2圖