



(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 I414161 B

(45)公告日：中華民國 102 (2013) 年 11 月 01 日

(21)申請案號：100103271

(22)申請日：中華民國 100 (2011) 年 01 月 28 日

(51)Int. Cl. : H04L12/24 (2006.01)

H04W28/08 (2009.01)

(71)申請人：國立交通大學(中華民國) NATIONAL CHIAO TUNG UNIVERSITY (TW)

新竹市大學路 1001 號

(72)發明人：廖德誠 LIAW, DER CHERNG (TW)

(74)代理人：陳昭誠

(56)參考文獻：

US 6,078,943

US 7,181,524B1

US 7,231,445B1

US 2009/0241176A1

審查人員：蔡鴻璟

申請專利範圍項數：9 項 圖式數：4 共 0 頁

(54)名稱

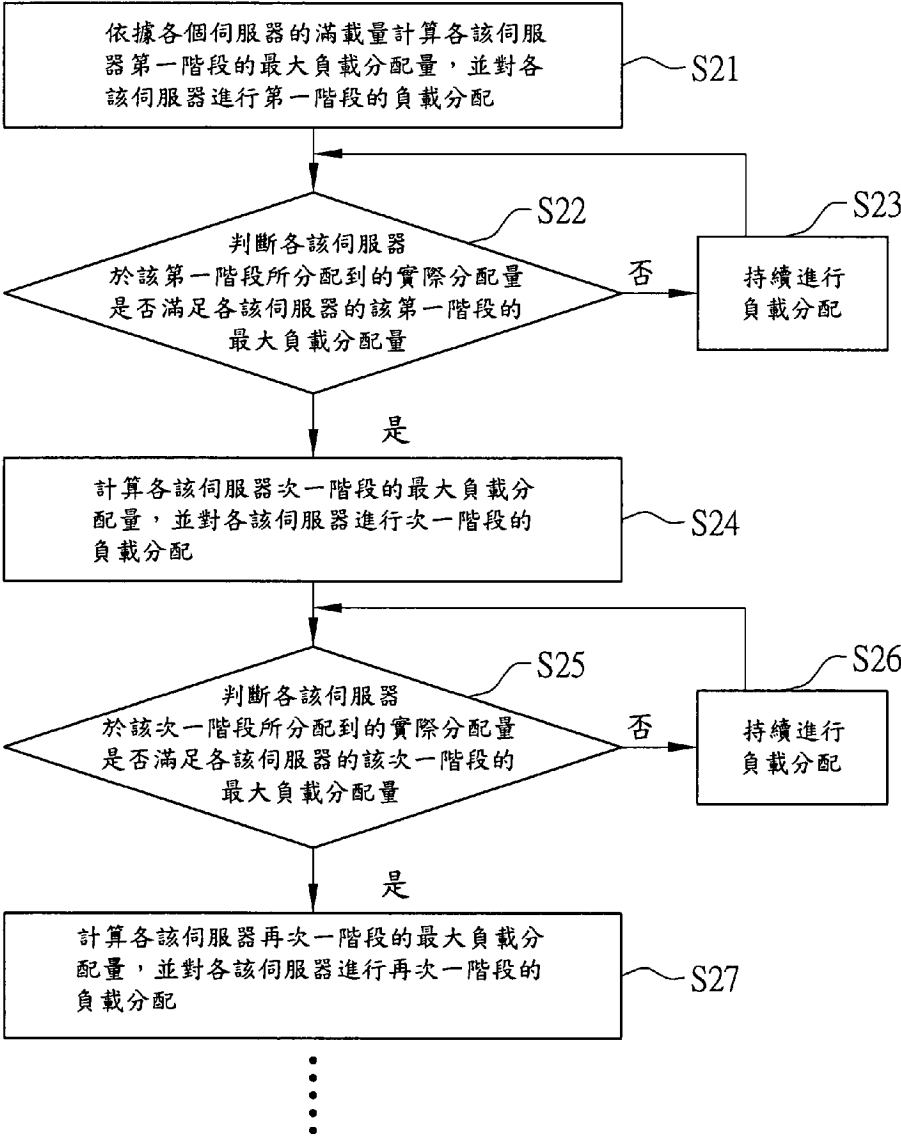
負載分配方法

METHOD FOR LOAD DISTRIBUTION

(57)摘要

一種負載分配方法，由負載分配裝置對複數個伺服器進行負載分配。先計算各伺服器第一階段的最大負載分配量並進行第一階段的負載分配。再判斷各伺服器於第一階段所分配到的實際分配量是否滿足各伺服器第一階段的最大負載分配量。於滿足時繼續計算各伺服器次一階段的最大負載分配量並判斷各伺服器於次一階段所分配到的實際分配量是否滿足各伺服器次一階段的最大負載分配量。持續判斷各伺服器於各階段中的實際分配量是否滿足各階段的最大負載分配量，其中，每一階段的最大負載分配量小於或等於其前一階段的最大負載分配量，且每一階段的最大負載分配量之累計總和小於或等於各伺服器的滿載量。據此，可依各階段的最大負載分配量進行負載分配以達到負載平衡。

This invention provides a method for a load distribution, wherein a load distribution device performs the load distribution to a plurality of servers. The method comprises steps of: firstly, calculating a max amount of the load distribution of a first stage and performing the load distribution of the first stage; then, determining whether an actual distributed amount of each of the plurality of servers in the first stage reaches the max amount of the load distribution of the first stage; if yes, subsequently calculating a max amount of the load distribution of a next stage and determining whether an actual distributed amount of each of the plurality of servers in the next stage reaches the max amount of the load distribution of the next stage; and continuously determining whether the actual distributed amount of each of the plurality of servers in each stage reaches the max amount of the load distribution of each stage, wherein the max amount of the load distribution of each stage is less than or equal to the max amount of the load distribution of a previous stage thereof, and an accumulated sum of the max amount of the load distribution of each stage is less than or equal to a full load amount of each server. Thereby, the load distribution can be performed in accordance with the max amount of the load distribution of each stage, so as to reach an load equilibrium.



第 2 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)

※申請案號：

(00103271)

※申請日：

100.1.28

※IPC 分類：

H04L 12/24 (2006.01)

H04W 58/08 (2009.01)

一、發明名稱：(中文/英文)

負載分配方法

METHOD FOR LOAD DISTRIBUTION

二、中文發明摘要：

一種負載分配方法，由負載分配裝置對複數個伺服器進行負載分配。先計算各伺服器第一階段的最大負載分配量並進行第一階段的負載分配。再判斷各伺服器於第一階段所分配到的實際分配量是否滿足各伺服器第一階段的最大負載分配量。於滿足時繼續計算各伺服器次一階段的最大負載分配量並判斷各伺服器於次一階段所分配到的實際分配量是否滿足各伺服器次一階段的最大負載分配量。持續判斷各伺服器於各階段中的實際分配量是否滿足各階段的最大負載分配量，其中，每一階段的最大負載分配量小於或等於其前一階段的最大負載分配量，且每一階段的最大負載分配量之累計總和小於或等於各伺服器的滿載量。據此，可依各階段的最大負載分配量進行負載分配以達到負載平衡。

三、英文發明摘要：

This invention provides a method for a load distribution, wherein a load distribution device performs the load distribution to a plurality of servers. The method comprises steps of: firstly, calculating a max amount of the load distribution of a first stage and performing the load distribution of the first stage; then, determining whether an actual distributed amount of each of the plurality of servers in the first stage reaches the max amount of the load distribution of the first stage; if yes, subsequently calculating a max amount of the load distribution of a next stage and determining whether an actual distributed amount of each of the plurality of servers in the next stage reaches the max amount of the load distribution of the next stage; and continuously determining whether the actual distributed amount of each of the plurality of servers in each stage reaches the max amount of the load distribution of each stage, wherein the max amount of the load distribution of each stage is less than or equal to the max amount of the load distribution of a previous stage thereof, and an accumulated sum of the max amount of the load distribution of each stage is less than or equal to a full load amount of each server. Thereby, the load distribution can be performed in accordance with the max amount of the load distribution of each stage, so as to reach an load equilibrium.

四、指定代表圖：

(一)本案指定代表圖為：第 (2) 圖。

(二)本代表圖之元件符號簡單說明：

S21~S27 步驟

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】

本發明係關於一種負載分配方法，詳而言之，係涉及一種處理網路系統中資訊流量之負載分配方法。

【先前技術】

隨著科技進步，網路已成為現代社會的重要技術。隨著網路人口逐漸增加，對於回應客戶端請求之伺服器的網路負載亦隨之增加。

以往的伺服器容易負載過大，例如，請參閱第 1 圖，其為習知技術之伺服器負載平衡方法的流程圖。於步驟 S11 中，設定每一個伺服器的流量警戒值，接著於步驟 S12 中，負載平衡器將所接收之客戶端請求平均分配至每一個伺服器。其次，於步驟 S13 中，當其中一個伺服器的負載達到其流量警戒值時，該伺服器發出一通知至該負載平衡器，以通知該負載平衡器莫再將客戶端請求分配至該伺服器，而應將客戶端請求分配至其餘的伺服器。

然而，上述將客戶端請求平均分配至每一個伺服器的機制，即使該客戶端請求的數量很少時仍須啟動所有的伺服器，如此會增加伺服端的耗能，此外，伺服器的流量警戒值為負載平衡器進行負載平衡時絕對不能超越的界線，因此若每一個伺服器的負載皆到達其流量警戒值時，客戶端一方便呈現完全無法連線的狀態，需排隊等其中之一連線結束才可能清出負載空間。另外，伺服器亦可能因為流量警戒值設定不良而造成過載、當機或斷線，使系統整體

效率下降。

故，如何提出一種負載分配方法，能讓複數個伺服器達到負載平衡，以防止伺服器過載，並紓解客戶端連線塞車的狀況，是為目前亟待解決的問題之一。

【發明內容】

為解決前述習知技術的問題，本發明提供一種負載分配方法，係應用於網路系統中，以對複數個伺服器進行負載分配。

本發明之負載分配方法，係由負載分配裝置對複數個伺服器進行負載分配，包括以下步驟：(1)依據各該伺服器的滿載量計算各該伺服器的第一階段的最大負載分配量，並進行第一階段的負載分配；(2)判斷各該伺服器於該第一階段所分配到的實際分配量是否滿足各該伺服器的該第一階段的最大負載分配量，若否，則返回步驟(1)後半部所述之負載分配步驟，持續對該複數個伺服器進行該第一階段的負載分配，若是，進至步驟(3)；(3)計算各該伺服器的次一階段的最大負載分配量，並進行該次一階段的負載分配；以及(4)判斷各該伺服器於該次一階段所分配到的實際分配量是否滿足各該伺服器的該次一階段的最大負載分配量，若否，則返回步驟(3)後半部所述之負載分配步驟，持續對該複數個伺服器進行該次一階段的負載分配，若是，則計算各該伺服器再次一階段的最大負載分配量，並對該複數個伺服器進行該再次一階段的負載分配，其中，各該伺服器的每一階段的最大負載分配量小於或等於其前一階

段的最大負載分配量，且各該伺服器的每一階段的最大負載分配量之累計總和小於或等於各該伺服器的滿載量。

於一實施形態中，各該伺服器的第一階段的最大負載分配量為各該伺服器的滿載量的 $\frac{1}{M}$ 倍，其中，M 為可變參數。此外，各該伺服器的滿載量係依據其 CPU 的運作能力、記憶體的空間大小、硬碟運轉狀態及/或伺服器負載連接數量而計算。

於另一實施形態中，各該伺服器的每一階段的最大負載分配量係為其前一階段的負載分配量的 $\frac{1}{N}$ 倍，其中，N 為可變參數。

於一較佳實施形態中，前述該 M 及 N 均為 2。

相較於習知技術，本發明之負載分配方法能避免習知技術中，當客戶端請求為少量時仍啟用多個伺服器而導致增加耗能的缺失，進而達成複數個伺服器負載平衡之功效。

【實施方式】

以下藉由特定的具體實施形態說明本發明之實施方式，熟悉此技術之人士可由本說明書所揭示之內容輕易地了解本發明之其他優點與功效，亦可藉由其他不同的具體實施形態加以施行或應用。

請參閱第 2 圖，本發明之負載分配方法包括步驟 S21 至 S27。

於步驟 S21 中，依據各個伺服器的滿載量計算各該伺

伺服器第一階段的最大負載分配量，並對各該伺服器進行第一階段的負載分配。詳言之，係由負載分配裝置來決定各該伺服器第一階段的最大負載分配量。於一實施形態中，伺服器的第一階段的最大負載分配量可為伺服器的滿載量的 $\frac{1}{M}$ 倍，其中， M 為可變參數。於一較佳實施形態中， M 可為 2。此外，伺服器的滿載量可依據其 CPU 的運作能力、記憶體의 剩餘空間大小、硬碟運轉狀態等其他因素來計算。接著進至步驟 S22。

於步驟 S22 中，判斷各該伺服器於該第一階段所分配到的實際分配量是否滿足各該伺服器的該第一階段的最大負載分配量。換言之，負載分配裝置於第一階段分配給各該伺服器的負載限制為伺服器的滿載量的 $\frac{1}{M}$ 倍，並判斷各該伺服器於該第一階段所分配到的實際分配量是否滿足各該伺服器的該第一階段的最大負載分配量。若否，則進至步驟 S23，若是，則進至步驟 S24。

於步驟 S23 中，持續進行負載分配。接著返回步驟 S22 以進行判斷。

於步驟 S24 中，計算各該伺服器次一階段的最大負載分配量，並對各該伺服器進行次一階段的負載分配，其中，該次一階段的最大負載分配量係小於或等於該第一階段的最大負載分配量，且該第一階段和該次一階段的最大負載分配量之累計總和係小於或等於伺服器的滿載量。於一實

施形態中，各該伺服器的第一階段的最大負載分配量為各該伺服器的滿載量的 $\frac{1}{M}$ 倍，該次一階段的最大負載分配量為第一階段的最大負載分配量的 $\frac{1}{N}$ 倍，其中，N 為可變參數，或 N 等於 M。接著進至步驟 S25。

於步驟 S25 中，判斷各該伺服器於該次一階段所分配到的實際分配量是否滿足各該伺服器的該次一階段的最大負載分配量。換言之，當 M 等於 N 時，負載分配裝置於該次一階段分配給各該伺服器的負載限制為伺服器的滿載量的 $\frac{1}{M^2}$ 倍，或者當 M 不等於 N 時為 $\frac{1}{MN}$ 倍，並判斷各該伺服器於該次一階段所分配到的實際分配量是否滿足各該伺服器於該次一階段的最大負載分配量。若否，則進至步驟 S26，若是，則進至步驟 S27。

於步驟 S26 中，持續進行負載分配。接著返回步驟 S25 以進行判斷。

於步驟 S27 中，計算各該伺服器之再次一階段的最大負載分配量，並對各該伺服器進行再次一階段的負載分配。以此類推，由於每一階段的最大負載分配量皆小於或等於其前一階段的最大負載分配量，且每一階段的最大負載分配量之累計總和係小於或等於伺服器的滿載量，因而各個伺服器的負載永遠不會超過其滿載。

另外，於進行負載分配時，該負載分配裝置可監測該

複數個伺服器，以記錄該複數個伺服器的每一階段的最大負載分配量及取得該複數個伺服器的每一階段的實際分配量。另外，該負載分配裝置亦可依據不同狀態(例如依據不同伺服器的硬體效能)個別設定或計算各該伺服器每一階段的最大負載分配量。

此外，上述步驟 S21、S23、S24、S26 及 S27 中進行負載分配的步驟係指該負載分配裝置依據所接收到的用戶端請求將負載平均分配、隨機分配或依不同狀態個別分配至該複數個伺服器。另外，該負載分配裝置依據所接收到的客戶端請求，可將負載依序分配或選擇分配至該複數個伺服器。再者，當某一階段中各該伺服器皆有分配到負載而該負載分配裝置繼續進行負載分配時，當其中一個伺服器的實際分配量下降至低於預定的判斷值時，令該負載分配裝置將所接收到的客戶端請求優先分配給該伺服器，其中，該預定的判斷值可為由負載分配裝置當初所決定之該階段的最大負載分配量。

接著以第 3 圖及第 4 圖說明本發明之一具體實施形態。第 3 圖為本發明之負載分配方法之具體實施示意圖。第 4 圖為本發明之負載分配方法之具體實施流程圖。

如第 3 圖所示，複數個客戶端 4 所發出的客戶端請求經網路傳輸至負載分配裝置 3，接著負載分配裝置 3 依據所接收的客戶端請求將對應的負載分配至各個伺服器 5，再由各伺服器 5 依據所分配到的負載提供相對應的服務至客戶端 4。負載分配裝置 3 將負載分配至複數個伺服器 5

的詳細流程如第 4 圖所示。

於步驟 S31 中，計算或設定各個伺服器第一階段的最大負載分配量(係為各個伺服器的滿載量的 $\frac{1}{M}$ 倍)，並依據所接收到的用戶端請求對各該伺服器進行第一階段的負載分配。接著進至步驟 S32。

於步驟 S32 中，判斷各該伺服器於該第一階段所分配到的實際分配量是否滿足各該伺服器的滿載量的 $\frac{1}{M}$ 倍。若否，則進至步驟 S33，若是，則進至步驟 S34。換言之，若各該伺服器於第一階段時的總負載(實際分配量)達到各該伺服器的滿載量的 $\frac{1}{M}$ 倍時，進至步驟 S34。

於步驟 S33 中，持續進行負載分配。接著返回步驟 S32。

於步驟 S34 中，計算或設定各該伺服器第二階段的最大負載分配量(係為伺服器的滿載量的 $\frac{1}{M^2}$ 倍，亦即令前述之可變參數 $N=M$ ，故第一階段的最大負載分配量 = $\frac{1}{M} \cdot (\text{滿載量})$ ，第二階段的最大負載分配量 = $\frac{1}{M} \cdot \frac{1}{M} \cdot (\text{滿載量}) = \frac{1}{M^2} \cdot (\text{滿載量})$ ，並依據所接收到的用戶端請求對各該伺服器進行第二階段的負載分配。接著進至步驟 S35。

於步驟 S35 中，判斷各該伺服器於該第二階段所分配到的實際分配量是否滿足各該伺服器的滿載量的 $\frac{1}{M^2}$ 倍。若否，則進至步驟 S36，若是，則進至步驟 S37。換言之，若各該伺服器於第二階段時的總負載達到各該伺服器的滿載量的 $\left(\frac{1}{M} + \frac{1}{M^2}\right)$ 倍時，進至步驟 S37。

於步驟 S36 中，持續進行負載分配。接著返回步驟 S35。

於步驟 S37 中，計算各該伺服器第三階段的最大負載分配量(係為伺服器的滿載量的 $\frac{1}{M^3}$ 倍)，並依據所接收到的用戶端請求對各該伺服器進行第三階段的負載分配。直到各該伺服器於第三階段時的總負載達到各該伺服器的滿載量的 $\left(\frac{1}{M} + \frac{1}{M^2} + \frac{1}{M^3}\right)$ 倍時，計算第四階段的最大負載分配量，以此類推。

由上述第 3 及 4 圖所述之具體實施形態可知，負載分配裝置將客戶端請求分配至複數個伺服器，其各階段的最大負載分配量係小於或等於其前一階段的最大負載分配量。例如，當 $N=M$ 時，各階段的最大負載分配量可依次為滿載量的 $\frac{1}{M}$ 、 $\frac{1}{M^2}$ 、 $\frac{1}{M^3}$ 、 $\dots$ 、 $\frac{1}{M^n}$ 倍，使得伺服器每一階段的最大負載分配量之累計總和依次為滿載量的 $\frac{1}{M}$ 、

$\left(\frac{1}{M} + \frac{1}{M^2}\right)$ 、 $\left(\frac{1}{M} + \frac{1}{M^2} + \frac{1}{M^3}\right)$ 、 $\dots$ 、 $\left(\frac{1}{M} + \frac{1}{M^2} + \frac{1}{M^3} + \dots + \frac{1}{M^n}\right)$ 倍。又例如，

當 $N=M=2$ 時，各階段的最大負載分配量之變數依次為 $\frac{1}{2}$ 、

$\frac{1}{4}$ 、 $\frac{1}{8}$ 、 $\dots$ 、 $\frac{1}{2^n}$ ；當 $N=M=3$ 時，各階段的最大負載分配量

之變數可依次 $\frac{1}{3}$ 、 $\frac{1}{9}$ 、 $\frac{1}{27} \dots \frac{1}{3^n}$ ，則每一階段的最大負載分配

量之累計總和不會超過伺服器的滿載量。

另外，當 $N \neq M$ 時，各階段的最大負載分配量之變數可依次為 $\frac{1}{M}$ 、 $\frac{1}{MN}$ 、 $\frac{1}{MN^2}$ 、 $\dots$ 、 $\frac{1}{MN^{n-1}}$ ，或者 $\frac{1}{M}$ 、 $\frac{1}{MN}$ 、 $\frac{1}{MN(N+1)}$ 、

$\frac{1}{MN(N+1)(N+2)}$ 、 $\dots$ 、 $\frac{1}{MN(N+1)\dots(N+n)}$ 等其他數學式的運算變

化。亦即，限制每一階段的最大負載分配量小於或等於其前一階段的最大負載分配量，且各階段的最大負載分配量之累計總和小於或等於伺服器的滿載量即可。如此一來，各伺服器的負載永遠不會超過自己的滿載。

綜上所述，本發明之負載分配方法，係計算各階段的最大負載分配量，並判斷伺服器於各該階段分配到的實際分配量是否滿足各該階段的最大負載分配量，且其中，每一階段的最大負載分配量係小於或等於其前一階段的最大負載分配量，且每一階段的最大負載分配量之累計總和係小於或等於滿載量。此外，該伺服器的第一階段的最大負

載分配量較佳為伺服器的滿載值的 $\frac{1}{M}$ 倍，M 係根據該伺服器的中央處理器（CPU）的運作能力、記憶體的空間大小、硬碟運轉狀態及/或伺服器負載連接數量或其他伺服器的相關資訊所計算之。因此，藉由本發明的運用，能使各個伺服器的負載平衡，更可提升客戶端連線效率。

上述實施形態僅例示性說明本發明之原理、特點及其功效，並非用以限制本發明之可實施範疇，任何熟習此項技藝之人士均可在不違背本發明之精神及範疇下，對上述實施形態進行修飾與改變。任何運用本發明所揭示內容而完成之等效改變及修飾，均仍應為下述之申請專利範圍所涵蓋。因此，本發明之權利保護範圍，應如後述之申請專利範圍所列。

【圖式簡單說明】

第 1 圖係習知技術之負載平衡方法之流程圖；

第 2 圖係本發明之負載分配方法之流程圖；

第 3 圖係本發明之負載分配方法之實施形態之示意圖；以及

第 4 圖係本發明之負載分配方法之一實施形態之流程圖。

【主要元件符號說明】

3 負載分配裝置

4 客戶端

5 伺服器

S11～S13、S21～S27、S31～S37 步驟

七、申請專利範圍：

1. 一種負載分配方法，係由負載分配裝置對複數個伺服器進行負載分配，包括以下步驟：

(1)依據各該伺服器的滿載量計算各該伺服器的第一階段的最大負載分配量，並進行該第一階段的負載分配；

(2)判斷各該伺服器於該第一階段所分配到的實際分配量是否滿足各該伺服器的該第一階段的最大負載分配量，若否，則返回步驟(1)後半部所述之負載分配步驟，持續對該複數個伺服器進行該第一階段的負載分配，若是，進至步驟(3)；

(3)計算各該伺服器的次一階段的最大負載分配量，並進行該次一階段的負載分配；以及

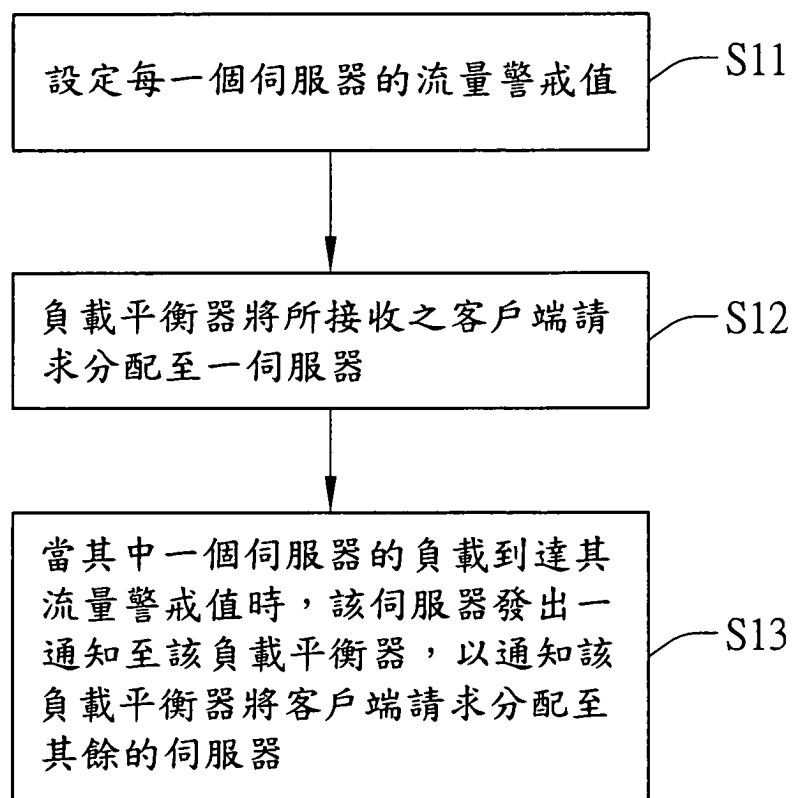
(4)判斷各該伺服器於該次一階段所分配到的實際分配量是否滿足各該伺服器的該次一階段的最大負載分配量，若否，則返回步驟(3)後半部所述之負載分配步驟，持續對該複數個伺服器進行該次一階段的負載分配，若是，則計算各該伺服器的再次一階段的最大負載分配量，並對該複數個伺服器進行該再次一階段的負載分配，其中，各該伺服器的每一階段的最大負載分配量小於或等於其前一階段的最大負載分配量，且各該伺服器的每一階段的最大負載分配量之累計總和小於或等於各該伺服器的滿載量。

2. 如申請專利範圍第 1 項所述之負載分配方法，其中，於

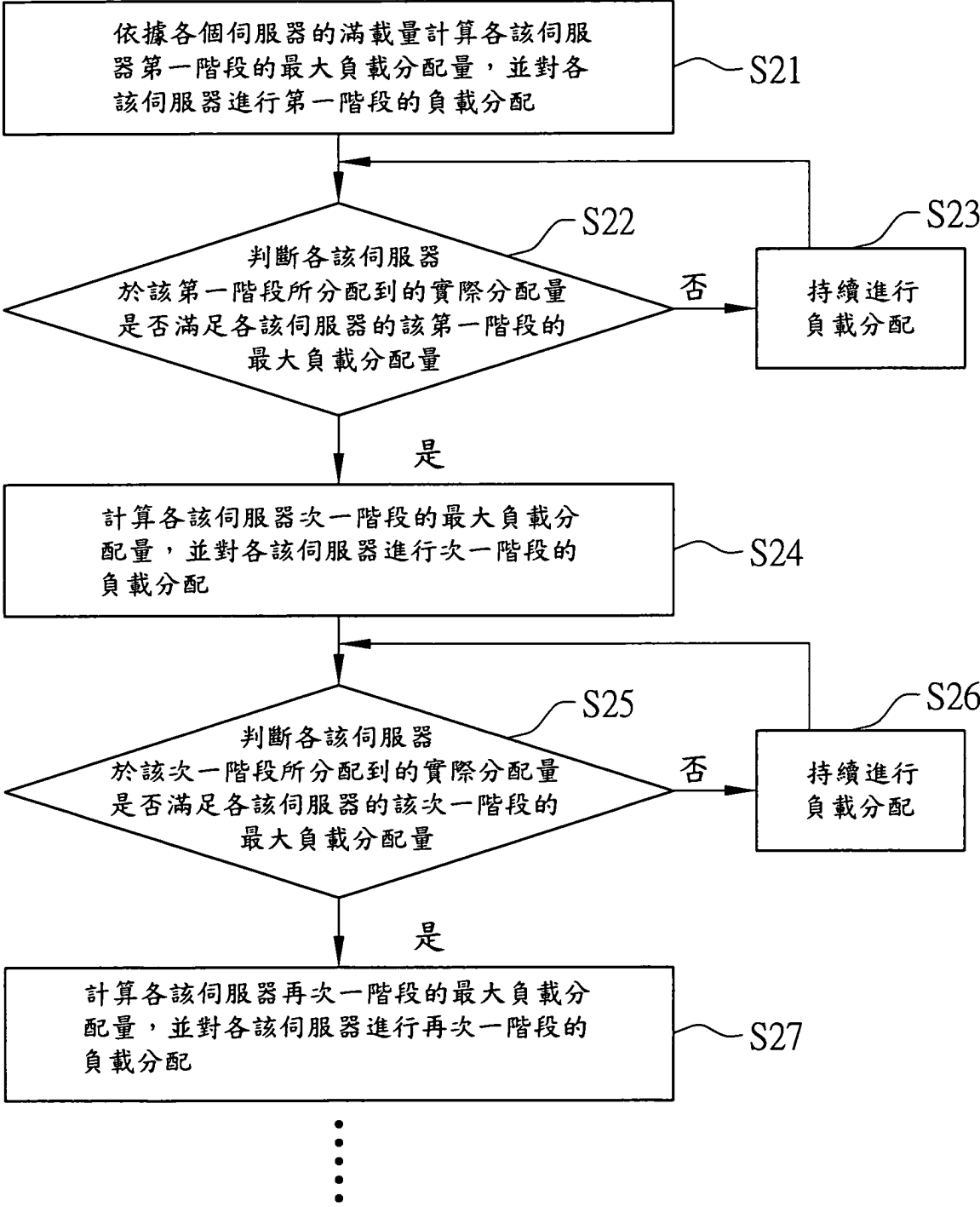
進行負載分配時，該負載分配裝置監測該複數個伺服器，以記錄該複數個伺服器的每一階段的最大負載分配量及取得該複數個伺服器的每一階段的實際分配量。

3. 如申請專利範圍第 1 項所述之負載分配方法，其中，進行負載分配的步驟係為該負載分配裝置依據所接收到的用戶端請求，依序將負載分配至該複數個伺服器。
4. 如申請專利範圍第 1 項所述之負載分配方法，其中，各該伺服器的滿載量，係依據各該伺服器的中央處理器的運作能力、記憶體的空間大小、硬碟運轉狀態或伺服器負載連接數量而計算。
5. 如申請專利範圍第 1 項所述之負載分配方法，其中，各該伺服器的第一階段的最大負載分配量為各該伺服器的滿載量的 $\frac{1}{M}$ 倍。
6. 如申請專利範圍第 5 項所述之負載分配方法，其中，M 為可變參數。
7. 如申請專利範圍第 5 項所述之負載分配方法，其中，各該伺服器的每一階段的最大負載分配量係為其前一階段的最大負載分配量的 $\frac{1}{N}$ 倍。
8. 如申請專利範圍第 7 項所述之負載分配方法，其中，N 為可變參數。
9. 如申請專利範圍第 1 項所述之負載分配方法，其中，該負載分配裝置於某一階段持續進行負載分配時，若已受

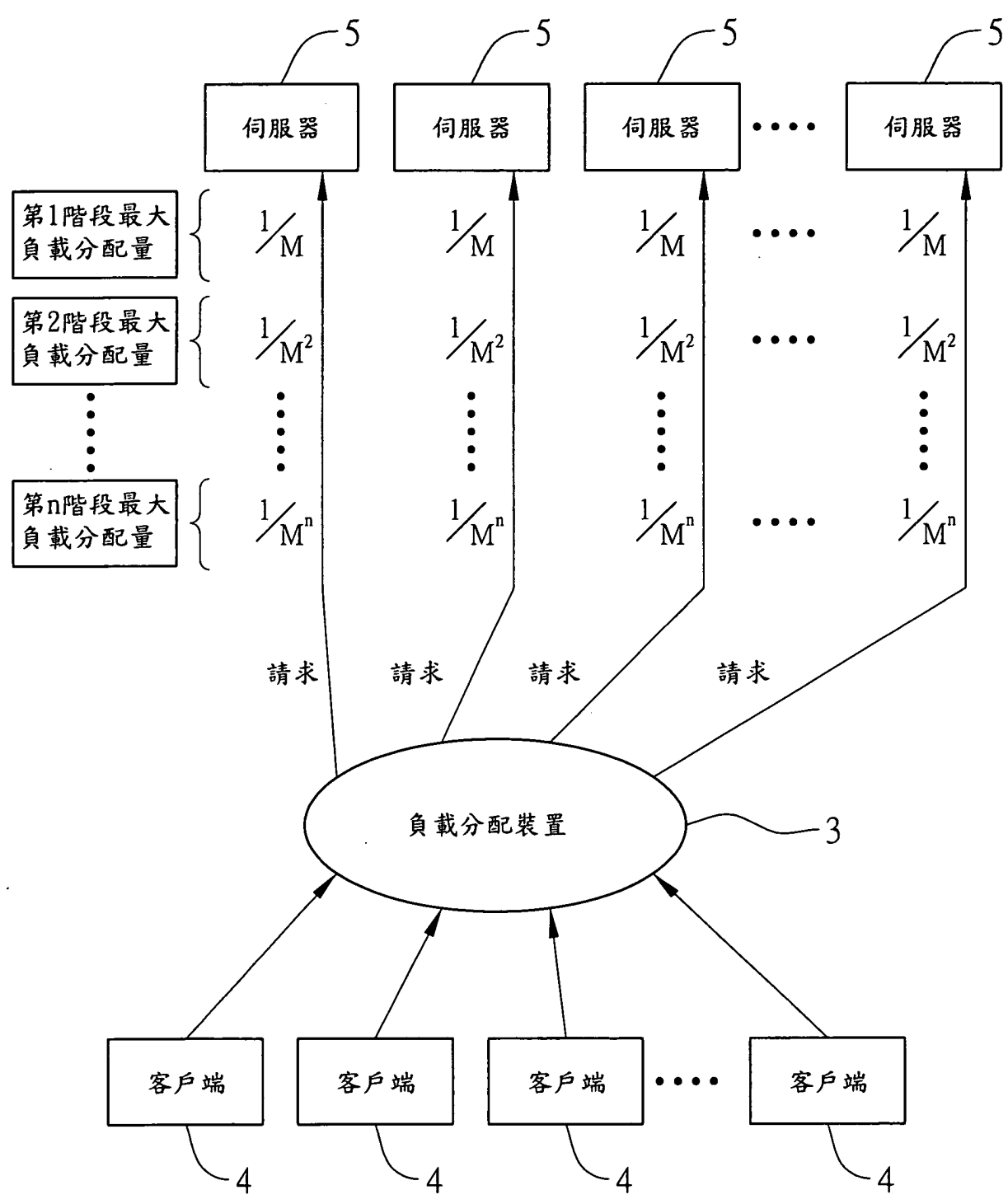
該階段負載分配之伺服器的實際分配量下降至低於預定的判斷值時，令該負載分配裝置將所接收到的客戶端請求優先分配給該伺服器。



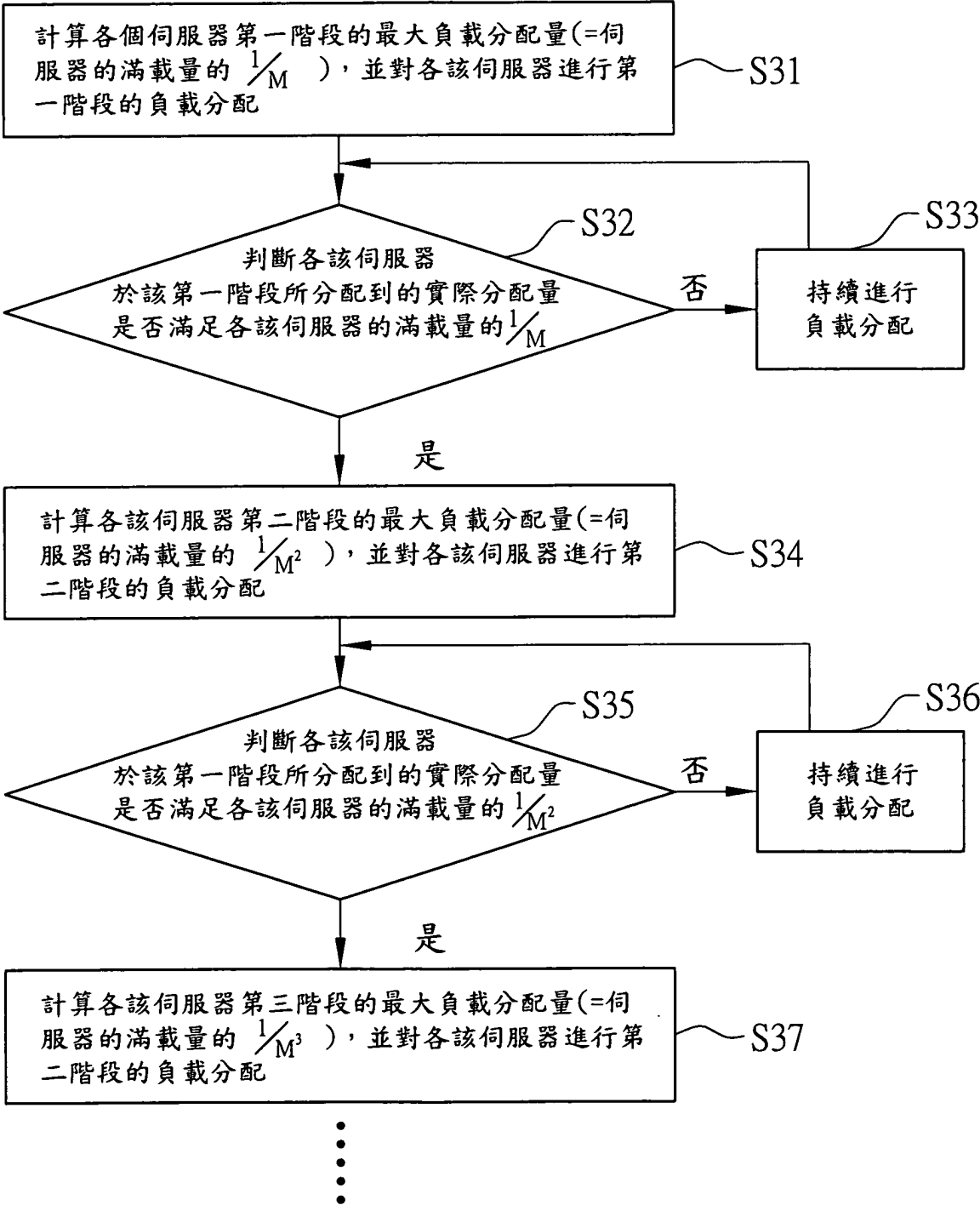
第 1 圖



第 2 圖



第 3 圖



第 4 圖