[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 9/455 (2006.01)
G06F 9/46 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200710199005.1

[43] 公开日 2008 年 4 月 30 日

[11] 公开号 CN 101169731A

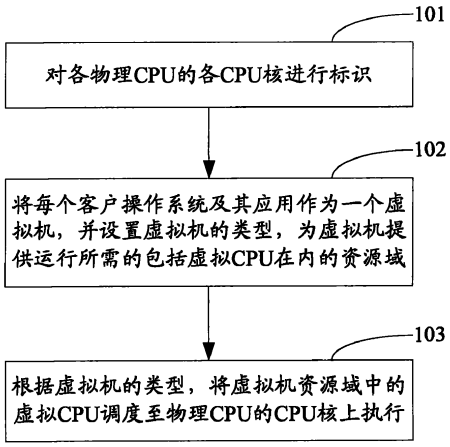
[22] 申请日 2007.12.5
[21] 申请号 200710199005.1
[71] 申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
[72] 发明人 翁楚良 全小飞
[74] 专利代理机构 北京德琦知识产权代理有限公司
代理人 宋志强 麻海明
权利要求书 5 页 说明书 11 页 附图 2 页

[54] 发明名称

多路多核服务器及其 CPU 的虚拟化处理方法

[57] 摘要

本发明公开了一种多路多核服务器的 CPU 虚拟化处理方法，包括：将每个客户操作系统及其应用作为一个虚拟机，并设置所述虚拟机的类型，为所述虚拟机提供运行所需的包括虚拟 CPU 在内的资源域；根据所述虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行。此外，本发明还公开了一种多路多核服务器。 本发明公开的技术方案，能够实现多路多核服务器的虚拟化，并且该虚拟化技术协同虚拟化硬件、物理硬件和操作系统三者之间关系，实现了优化的性能和效率。



1、一种多路多核服务器的 CPU 虚拟化处理方法，其特征在于，该方法包括:

将每个客户操作系统及其应用作为一个虚拟机，并设置所述虚拟机的类型，根据所述虚拟机的类型为所述虚拟机提供运行所需的包括虚拟 CPU 在内的资源域;

根据所述虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行。

2、如权利要求 1 所述的方法，其特征在于，所述虚拟机的类型包括：并发型和/或吞吐型。

3、如权利要求 2 所述的方法，其特征在于，所述虚拟机类型为并发型时，所述资源域内的虚拟 CPU 的个数小于等于所有物理 CPU 的 CPU 核的个数。

4、如权利要求 2 所述的方法，其特征在于，所述根据虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行包括:

根据虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 加入到物理 CPU 的 CPU 核的运行队列中，在触发虚拟 CPU 调度时，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行。

5、如权利要求 4 所述的方法，其特征在于，所述触发虚拟 CPU 调度为：发生时间片轮转或虚拟 CPU 阻塞或虚拟 CPU 唤醒或中断。

6、如权利要求 4 所述的方法，其特征在于，所述根据虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 加入到物理 CPU 的 CPU 核的运行队列中包括:

根据虚拟机的类型，确定虚拟机资源域中的虚拟 CPU 可以加入的 CPU 核的集合;

从所述确定的 CPU 核集合中，选取负载最轻的 CPU 核，将所述虚拟 CPU 加入至所选取的 CPU 核的运行队列中。

7、如权利要求 6 所述的方法，其特征在于，当前虚拟 CPU 所属虚拟机的

类型为并发型，所述虚拟 CPU 可以加入的 CPU 核的集合为：运行队列中没有与当前虚拟 CPU 属于同一虚拟机的虚拟 CPU 的 CPU 核的集合；

其中，所述 CPU 核的相邻核上已分配了与当前虚拟 CPU 同属一个虚拟机的虚拟 CPU；或者，与当前虚拟 CPU 同属一个虚拟机的虚拟 CPU 已恰好分配满了一个或多个 CPU 核时或尚未分配时，所述 CPU 核的相邻核上没有分配与当前虚拟 CPU 同属一个虚拟机中的虚拟 CPU。

8、如权利要求 6 所述的方法，其特征在于，当前虚拟 CPU 所属虚拟机的类型为吞吐型，所述虚拟 CPU 可以加入的 CPU 核的集合为：所有可用 CPU 核的集合。

9、如权利要求 4 所述的方法，其特征在于，该方法进一步包括：预先为虚拟机的资源域设置占用所有物理 CPU 使用率的权重及资源域内各虚拟 CPU 的势能初值；

将所述虚拟机资源域中的虚拟 CPU 加入到 CPU 核的运行队列中之后，进一步包括：达到预设时间间隔时，根据所述权重及资源域内的虚拟 CPU 数量，更新所述运行队列中的所述虚拟 CPU 的势能，按照所述虚拟 CPU 的势能，对所述虚拟 CPU 进行降序排列；

所述将运行队列中的虚拟 CPU 映射至 CPU 核上执行为：根据虚拟机的类型及所述运行队列头部的虚拟 CPU 势能，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行。

10、如权利要求 9 所述的方法，其特征在于，所述根据权重及资源域内的虚拟 CPU 数量，更新所述运行队列中的所述虚拟 CPU 的势能为：

根据 CPU 核的总个数、预设时间间隔内的滴答次数及每次滴答的势能消耗，计算得到总势能；

根据所述总势能、权重及资源域内的虚拟 CPU 数量，计算得到资源域内每个虚拟 CPU 能量增量；

根据运行队列中的虚拟 CPU 的原势能及所述能量增量，计算得到所述虚拟 CPU 的更新后的势能。

11、如权利要求 9 所述的方法，其特征在于，所述根据虚拟机的类型及所述运行队列头部的虚拟 CPU 势能，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行包括:

当前 CPU 核的运行队列头部的虚拟 CPU 势能小于零，且所述虚拟 CPU 所属的虚拟机类型为并发型，则寻找其它 CPU 核运行队列头部的势能大于零的虚拟 CPU，并从中选取势能值最大且与当前 CPU 核运行队列中的虚拟 CPU 不属于同一虚拟机的虚拟 CPU，将选定的运行队列头部的虚拟 CPU 映射至当前 CPU 核上运行。

12、如权利要求 9 所述的方法，其特征在于，所述根据虚拟机的类型及所述运行队列头部的虚拟 CPU 势能，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行包括:

当前 CPU 核的运行队列头部的虚拟 CPU 势能小于零，且所述虚拟 CPU 所属的虚拟机类型为吞吐型，则寻找其它 CPU 核运行队列头部的势能大于零的虚拟 CPU，并从中选取势能值最大的虚拟 CPU，将选定的运行队列头部的虚拟 CPU 映射至当前 CPU 核上运行。

13、如权利要求 9 所述的方法，其特征在于，所述根据虚拟机的类型及所述运行队列头部的虚拟 CPU 势能，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行包括:

当前 CPU 核的运行队列头部的虚拟 CPU 势能大于零，且所述虚拟 CPU 所属的虚拟机类型为并发型，则向同属该虚拟机的虚拟 CPU 所在的 CPU 核发送处理器间中断 IPI，并将所述虚拟 CPU 映射至当前的 CPU 核上运行；接收到所述 IPI 后的 CPU 核将同属该虚拟机的虚拟 CPU 从运行队列中取出并插入到运行队列的头部，然后发送调度软中断触发调度。

14、如权利要求 9 所述的方法，其特征在于，所述根据虚拟机的类型及所述运行队列头部的虚拟 CPU 势能，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行包括:

当前 CPU 核的运行队列头部的虚拟 CPU 势能大于零，且所述虚拟 CPU 所

属的虚拟机类型为吞吐型，则直接将所述虚拟 CPU 映射至 CPU 核上运行。

15、如权利要求 1 所述的方法，其特征在于，该方法进一步包括：对各物理 CPU 的各 CPU 核进行标识，根据所述标识识别所述虚拟 CPU 被调度到的物理 CPU 的 CPU 核。

16、如权利要求 15 所述的方法，其特征在于，所述资源域内的每个虚拟 CPU 为一个数据结构，所述数据结构包括：CPU 核标识属性；

所述将虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上之后，进一步包括：将所述 CPU 核标识属性赋值为所述物理 CPU 的 CPU 核的标识。

17、如权利要求 15 所述的方法，其特征在于，所述对各物理 CPU 的各 CPU 核进行标识为：采用线性方式对各物理 CPU 的各 CPU 核进行标识；或者为：采用矩阵方式对各物理 CPU 的各 CPU 核进行标识。

18、如权利要求 17 所述的方法，其特征在于，所述采用矩阵方式对各物理 CPU 的各 CPU 核进行标识包括：

为各物理 CPU 的各 CPU 核设置线性编号标识；

对所述线性编号进行转换处理，得到所述 CPU 核对应的物理 CPU 编号和核编号；

利用预设的长度为 2N 位的无符号整型数的高 N 位表示所述物理 CPU 编号，低 N 位表示所述核编号。

19、一种多路多核服务器，其特征在于，该多路多核服务器包括：

多个物理 CPU，每个物理 CPU 包括多个 CPU 核；

虚拟机监控器，用于将每个客户操作系统及其应用作为一个虚拟机，并设置所述虚拟机的类型，为所述虚拟机提供运行所需的包括虚拟 CPU 在内的资源域；根据所述虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行。

20、如权利要求 19 所述的多路多核服务器，其特征在于，所述虚拟机监控器包括：

资源域设置模块，用于将每个客户操作系统及其应用作为一个虚拟机，并

设置所述虚拟机的类型，为所述虚拟机提供运行所需的包括虚拟 CPU 在内的资源域;

CPU 调度模块，用于根据所述虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行。

21、如权利要求 20 所述的多路多核服务器，其特征在于，所述 CPU 调度模块包括:

运行队列加入模块，用于根据虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 加入到物理 CPU 的 CPU 核的运行队列中;

调度模块，用于在触发虚拟 CPU 调度时，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行。

22、如权利要求 21 所述的多路多核服务器，其特征在于，所述运行队列加入模块包括:

集合确定模块，用于根据虚拟机的类型，确定虚拟机资源域中的虚拟 CPU 可以加入的 CPU 核的集合;

加入模块，用于从所述确定的 CPU 核集合中，选取负载最轻的 CPU 核，将所述虚拟 CPU 加入至所选取的 CPU 核的运行队列中。

23、如权利要求 21 所述的多路多核服务器，其特征在于，该多路多核服务器进一步包括:

虚拟 CPU 势能计算模块，用于在达到预设时间间隔时，根据预先设置的虚拟机的资源域占用所有物理 CPU 使用率的权重及资源域内的虚拟 CPU 数量，计算运行队列中的虚拟 CPU 的势能;

虚拟 CPU 排序模块，用于按照运行队列中虚拟 CPU 的势能，对所述虚拟 CPU 进行降序排列;

所述调度模块在触发虚拟 CPU 调度时，根据虚拟机的类型及运行队列头部的虚拟 CPU 势能，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行。

多路多核服务器及其CPU的虚拟化处理方法

技术领域

本发明涉及计算机技术，尤其涉及一种多路多核服务器及其中央处理器（CPU）的虚拟化处理方法。

背景技术

多核处理器（多核CPU）系统也即单片多处理器（Chip MultiProcessor，CMP）系统，是指由单个芯片上的多个处理器核所构成的多处理器系统。CMP 允许线程在多个处理器核上并行执行，从而利用线程级并行提高系统性能。多路多核服务器是指包括多个多核处理器芯片的计算机服务系统。

系统级虚拟化技术通常是在计算机硬件和操作系统之间增加虚拟机监控器(Virtual Machine Monitor，VMM)，通过虚拟机监控器向上层操作系统提供下层硬件上的虚拟化，以解除二者间的直接依赖。随着x86体系结构处理器等通用处理器性能的提高，由于虚拟化技术可以有效降低成本、易管理、提高系统可用性、动态负载平衡、加强安全策略等特点，使得微处理器计算系统虚拟化技术成为目前的技术新动向。

传统的系统级虚拟化技术基于全虚拟化原理（full-virtualization），将传统的直接执行和快速的动态二进制翻译技术结合起来，即由虚拟机监控器向上层操作系统提供整个下层硬件上的虚拟化，采用二进制翻译技术，但由于虚拟机监控器模拟整个硬件环境，因此存在较大的系统开销，导致虚拟化实现效率低。

为此，提出一种基于半虚拟化原理（para-virtualization）的虚拟化技术，即部分虚拟化下层硬件环境，同时修改上层操作系统的部分功能，以实现多个虚拟机同时运行在宿主机上，该方法协同虚拟化硬件、物理硬件和操作系

统三者之间关系，以实现优化的性能和效率，但该方案只针对单核处理器提出了解决方案，针对多路多核服务器尚没有具体解决方案。

发明内容

有鉴于此，本发明实施例中一方面提供一种多路多核服务器的 CPU 虚拟化处理方法，另一方面提供一种多路多核服务器，以便实现多路多核服务器的虚拟化。

本发明实施例提供的多路多核服务器的 CPU 虚拟化处理方法，包括：

将每个客户操作系统及其应用作为一个虚拟机，并设置所述虚拟机的类型，为所述虚拟机提供运行所需的包括虚拟 CPU 在内的资源域；

根据所述虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行。

本发明实施例提供的多路多核服务器，包括：

多个物理 CPU，每个物理 CPU 包括多个 CPU 核；

虚拟机监控器，用于将每个客户操作系统及其应用作为一个虚拟机，并设置所述虚拟机的类型，为所述虚拟机提供运行所需的包括虚拟 CPU 在内的资源域；根据所述虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行。

从上述方案可以看出，本发明实施例中通过将每个客户操作系统及其应用作为一个虚拟机，并设置所述虚拟机的类型，为所述虚拟机提供运行所需的包括虚拟 CPU 在内的资源域（即虚拟化硬件），然后根据虚拟机的类型，将虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行，从而实现多路多核服务器的虚拟化，该虚拟化技术协同虚拟化硬件、物理硬件和操作系统三者之间关系，实现了优化的性能和效率。

附图说明

图 1 为本发明实施例中多路多核服务器的 CPU 虚拟化处理方法的示例

性流程图；

图 2 为本发明实施例中多路多核服务器的示例性结构图；

图 3 为图 2 所示多路多核服务器中虚拟机监控器的结构示意图。

具体实施方式

本发明实施例中，基于半虚拟化原理，提出多路多核服务器的 CPU 虚拟化处理方案，针对多核处理器的特点、以及处理器核数不断增长的趋势，结合服务器应用运行情况，通过将每个客户操作系统及其应用作为一个虚拟机，并设置所述虚拟机的类型，为所述虚拟机提供运行所需的包括虚拟 CPU 在内的资源域；根据所述虚拟机的类型，将所述虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行。为了识别每个 CPU 核，可对 CPU 核进行标识，根据所述标识为各 CPU 核调度虚拟 CPU 并执行。

下面结合实施例和附图，进一步详细说明。

图 1 为本发明实施例中多路多核服务器的 CPU 虚拟化处理方法的示例性流程图。如图 1 所示，该流程包括如下步骤：

步骤 101，对各物理 CPU 的各 CPU 核进行标识。

通常情况下，操作系统将一个有 n 核的物理 CPU 看作对等的 n 个 CPU，多个 CPU 动态地从系统的就绪进程队列中调度任务并加以执行，一个进程在不同的时期可以在不同的 CPU 上运行，中断请求动态地在多个 CPU 间进行分配，并由指派的 CPU 提供中断服务。

为了标识物理 CPU 的各个核，通常情况下，采用线性方式标识系统中的 CPU 核，例如，对于 m 个物理 CPU，每个物理 CPU 包括 n 个 CPU 核的情况，用线性方式表示时，可对所有 CPU 核进行统一编号，得到线性编号：0, 1, 2, …, m × n-1。

本实施例中，针对多路多核的特性，以矩阵的方式标识系统中的多核 CPU。例如，cpu(0,0)、cpu(0,1)、cpu(1,0)、cpu(1,1)即表示系统中有两个物理 CPU，在每个物理 CPU 中分别有两个核。同样，对于 m 个物理 CPU，每

个物理 CPU 包括 n 个 CPU 核的情况，CPU 核的编号可以为：(0,0), (0, 1), …, (0, n-1), …, (m-1, 0), …, (m-1, n-1)。以矩阵的方式标识系统中的多核 CPU，可以有效反映 CPU 核间的关系。

在具体实现上，可设置长度为 2N 位的无符号整型数，用该整型数的高 N 位表示物理 CPU 编号，低 N 位表示物理 CPU 内的核编号。

此外，为了与现有技术兼容，可承继现有的虚拟机监控器采用的线性表示方法标识 CPU 核，例如 CPU 核的线性标识由整型数“processor”表示，并设置函数“smp_process_id()”，用于获得当前进程所在的 CPU 核的线性编号。当需要获取该 CPU 核的详细信息时，可通过预设的映射函数实现从线性表示方式向矩阵表示方式的转换。例如，针对同构 CPU 的服务器系统，可预先设置映射函数“get_coreid_from_processor()”，该函数中设置算法包括：物理 CPU 的编号由 processor / n 得到，物理 CPU 内的核编号由 processor mod n 得到。标识该 CPU 核时，可由一个无符号长整型记录，例如可设置 64 位的无符号长整型数“processorL”，其中，高 32 位用于表示 CPU 核所在的物理 CPU 编号，低 32 位用于表示 CPU 核在物理 CPU 内的核编号。

这样，在虚拟机监控器中，可通过 get_coreid_from_processor(smp_process_id())即可得到服务器当前进程所在的 CPU 核的标识。

步骤 102，将每个客户操作系统及其应用作为一个虚拟机，并设置虚拟机的类型，为虚拟机提供运行所需的包括虚拟 CPU 在内的资源域。

本实施例中，将每个运行于虚拟机监控器之上的客户操作系统及其应用，称为一个运行于虚拟机监控器上的虚拟机，为虚拟机提供运行所需的包括虚拟 CPU 在内的资源域，每个虚拟机分别运行在一个资源域中。资源域是下层硬件物理资源虚拟化后，提供给上层客户操作系统运行的平台，或者说是客户操作系统所见到的“硬件资源”。资源域中，包括虚拟 CPU（vCPU）、虚拟内存和虚拟 I/O 等。其中，vCPU 可以由一个数据结构描述其属性，数据结构中除了包含计时器、调度数据、寄存器信息、内存页表基址信息等，

针对多核的特性，还包括对应的物理 CPU 核标识，用以在调度至物理 CPU 的 CPU 核上时标识该 CPU 核，即将 CPU 核标识属性赋值为所调度到的 CPU 核标识，以表示其被动态映射至某一 CPU 中的某核上。

此外，本实施例中，根据服务器应用的特点，将虚拟机的类型分为吞吐（high-throughput）型虚拟机和并发（concurrent）型虚拟机两类。其中，吞吐型虚拟机上的应用一般为多个线程或进程，进程或线程之间不存在同步操作。例如，Web 服务器应用，每接收到一个访问请求，即刻动态生成一个线程用以响应用户的请求，线程与线程之间没有同步操作。并发型虚拟机上的应用一般是并行执行的进程或线程，在进程或线程之间的同步操作频繁。例如，科学计算中的消息通信接口（Message Passing Interface，MPI）并行进程程序或开放多线程（OpenMP）并行线程程序，在进程或线程之间需要进行频繁的同步操作（如进程之间需要交换数据）。相应地，在资源域的属性中包含有其上运行的虚拟机类型，虚拟机类型由用户在创建虚拟机时指定。其中，由系统启动的控制虚拟机（即控制虚拟机监控器的虚拟机）则由系统默认设定为吞吐型。

当虚拟机类型为并发型时，为了实现同步操作，资源域内的 vCPU 的个数小于等于所有物理 CPU 的 CPU 核的总个数，而当虚拟机类型为吞吐型时，资源域内的 vCPU 的个数不受限制。

具体实现时，资源域可由一个数据结构 struct domain 定义。根据服务器应用的特点，虚拟机监控器在资源域上构建两种类型的虚拟机，即吞吐型和并发型虚拟机，它们在 vCPU 调度策略上不同。除此之外，它们的其它属性相同。为了定义虚拟机的类型，在资源域对应的数据结构 struct domain 中相关属性如下：

```
struct domain                    /* 域结构 */
{
  domtype   domain_type;         /* 类型 */
  domid_t   domain_id;           /*  ID 号 */
```

```
    shared_info_t    *shared_info;  /* 共享信息 */
    spinlock_t     big_lock;
    …
}
```

其中，domtype 是枚举类型，其定义如下：

```
typedef enum
{HIGH_THROUGH,      /* 吞吐虚拟机类型 */
 CONCURRENCE        /* 并发虚拟机类型 */
} domtype;
```

此外，vCPU 的相关属性可如下所示：

```
struct vcpu
{
   unsigned long processorL;     /* CPU 核标识属性 */
   int vcpu_id;
    vcpu_info_t   *vcpu_info;
   struct domain     *domain;
  …
}
```

其中，变量 processorL 若为 64 位，则其高 32 位用于表示其对应物理 CPU 核的 CPU 编号，低 32 位用于表示其对应物理 CPU 核在处理器内的核编号。

步骤 103，根据虚拟机的类型，将虚拟机资源域中的虚拟 CPU 调度至物理 CPU 的 CPU 核上执行。

本实施例中，对于吞吐型虚拟机，虚拟机监控器在调度时，以 vCPU 为单位进行调度，即动态的将就绪的 vCPU 调度至空闲的物理 CPU 核上，以最大化虚拟机处理作业的吞吐量。对于并发型虚拟机，虚拟机监控器在调度时，以资源域为单位实现调度，即虚拟机监控器协同调度同一个虚拟机中的多个 vCPU，而不是独立调度多个虚拟机中的不同 vCPU。

实际应用中，每个物理 CPU 核维护一个运行队列（runq），对于每个

活动的 vCPU,在其生成或需要迁移时需要将其加入至某一物理 CPU 的 CPU 核的 runq 队列中，本实施例中，可根据虚拟机的类型，将虚拟机资源域中的 vCPU 加入到物理 CPU 的 CPU 核的运行队列中，之后在触发虚拟 CPU 调度时，将运行队列中的 vCPU 映射至 CPU 核上执行。

其中，根据虚拟机的类型，将虚拟机资源域中的 vCPU 加入到物理 CPU 的 CPU 核的运行队列中时，可首先根据虚拟机的类型，确定虚拟机资源域中的 vCPU 可以加入的 CPU 核的集合，然后从所确定的 CPU 核集合中，选取负载最轻的 CPU 核，将 vCPU 加入至所选取的 CPU 核的运行队列中。

例如：对属于并发型虚拟机的 vCPU，其可以加入的 CPU 核的集合中，每个 CPU 核的运行队列中没有来自同一虚拟机的 vCPU，即其可以加入的 CPU 核的集合为运行队列中没有来自同一虚拟机的 vCPU（即与当前 vCPU 属于同一虚拟机的 vCPU）的 CPU 核的集合。此外，该集合中的每个 CPU 核的相邻核上应该已分配了与当前 vCPU 同属一个虚拟机的虚拟 CPU；或者，当来自同一个虚拟机的 vCPU 已恰好分配满了一个或多个 CPU 核时或尚未分配时，该集合中的 CPU 核的相邻核上可以没有分配来自同一个虚拟机中的 vCPU。对属于吞吐型虚拟机的 vCPU，其可以加入的 CPU 核的集合为所有可用 CPU 核的集合。

确定 vCPU 可以加入的 CPU 核的集合后，从所确定的 CPU 核集合中，选取负载最轻的 CPU 核，将 vCPU 加入至所选取的 CPU 核的运行队列中。

通常情况下，vCPU 向物理 CPU 核的调度，是以时间片轮转的方式实现的，即一个 vCPU 调度到物理 CPU 核上，将会运行一个固定的时间（通常称一个滴答），同时，为了确定当前需要调度的 vCPU，可为每个 vCPU 设置势能参数，并且定期更新势能参数的取值，以便根据每个 vCPU 的势能值从运行队列中选取 vCPU 并映射至物理 CPU 核上。

具体实现时，预先为虚拟机的资源域设置权重及资源域内各 vCPU 的势能初值，权重即是每一资源域占用系统总体物理 CPU 使用率的百分比。其中，势能初值可取零值。

对于每个 CPU 核运行队列中的 vCPU，若 CPU 核为引导系统的处理器（bootstrap processor，BSP），则该 CPU 核按照预设时间间隔Δt，根据所设置的权重及资源域内的 vCPU 数量，更新运行队列中的 vCPU 的势能，之后对运行队列中的 vCPU 进行降序排列，其它 CPU 核则只根据各 vCPU 的势能对 vCPU 进行降序排列。对于每个 CPU 核上运行的 vCPU，在每个滴答后，该 vCPU 将消耗一定的势能。

其中，运行队列中的 vCPU 的势能更新过程可以包括:

A、根据 CPU 核的总个数、预设时间间隔内的滴答次数及每次滴答的势能消耗，计算得到总势能，即总势能 = 物理 CPU 核总数 × 每次滴答的势能消耗 × Δt 内的滴答次数。

B、根据总势能、权重及资源域内的 vCPU 数量，计算得到资源域内每个 vCPU 能量增量，即能量增量 = 总势能 × 权重 ÷ 该资源域中 vCPU 数。

C、根据运行队列中的 vCPU 的原势能及上述能量增量，计算得到 vCPU 的更新后的势能，即 vCPU 的势能 =原势能值 + 能量增量。

CPU 核上运行的 vCPU 的势能更新过程可以为：vCPU 势能 = 原势能值 - 每次滴答的势能消耗。

之后，在触发 vCPU 调度时，即发生时间片轮转或 vCPU 阻塞或 vCPU 唤醒或中断等情况时，根据虚拟机的类型及运行队列头部的 vCPU 势能，将运行队列中的 vCPU 映射至 CPU 核上执行。

具体映射过程包括:

若当前 CPU 核的运行队列头部的 vCPU 势能小于零，且该 vCPU 所属的虚拟机类型为并发型，则寻找其它 CPU 核运行队列头部的势能大于零的 vCPU，并从中选取势能值最大且与当前 CPU 核运行队列中的 vCPU 不属于同一虚拟机的 vCPU，将选定的运行队列头部的 vCPU 映射至当前 CPU 核上运行。

若当前 CPU 核的运行队列头部的 vCPU 势能小于零，且该 vCPU 所属的虚拟机类型为吞吐型，则寻找其它 CPU 核运行队列头部的势能大于零的

vCPU，并从中选取势能值最大的 vCPU，将选定的运行队列头部的 vCPU 映射至当前 CPU 核上运行。

若当前 CPU 核的运行队列头部的 vCPU 势能大于零，且该 vCPU 所属的虚拟机类型为并发型，则向同属该虚拟机的 vCPU 所在的 CPU 核发送处理器间中断（InterProcessor Interrupt，IPI），并将所述 vCPU 映射至当前的 CPU 核上运行；接收到所述 IPI 后的 CPU 核将同属该虚拟机的 vCPU 从运行队列中取出并插入到运行队列的头部，然后发送调度软中断触发调度。

若当前 CPU 核的运行队列头部的 vCPU 势能大于零，且该 vCPU 所属的虚拟机类型为吞吐型，则直接将所述 vCPU 映射至 CPU 核上运行。

以上对本发明实施例中的多路多核服务器的 CPU 虚拟化处理方法进行了详细描述，下面再对本发明实施例中的多路多核服务器进行详细描述。

图 2 是出了本发明实施例中的多路多核服务器的示例性结构图。如图 2 所示，该多路多核服务器包括：多个物理 CPU（图中示出了 2 个物理 CPU，实际应用中物理 CPU 的个数还可以是其它值），每个物理 CPU 包括多个 CPU 核（图中示出了 2 个 CPU 核，实际应用中 CPU 核的个数还可以是其它值）。此外，该多路多核服务器还包括一个虚拟机监控器，用于将每个客户操作系统及其应用作为一个虚拟机，并设置虚拟机的类型，为虚拟机提供运行所需的包括 vCPU 在内的资源域；之后，根据虚拟机的类型，将虚拟机资源域中的 vCPU 调度至物理 CPU 的 CPU 核上执行。此外，为了识别每个 CPU 核，该多路多核服务器还可对 CPU 核进行标识，根据所述标识为各 CPU 核调度虚拟 CPU 并执行。

其中，图 2 所示的多路多核服务器的具体操作过程可与图 1 所示方法流程中的操作过程一致。

具体实现时，虚拟机监控器的内部结构可有多种具体实现形式，图 3 示出了其中一种结构示意图。如图 3 所示，该虚拟机监控器可包括：资源域设置模块和 CPU 调度模块。

其中，资源域设置模块用于将每个客户操作系统及其应用作为一个虚拟机，

并设置所述虚拟机的类型，为所述虚拟机提供运行所需的包括 vCPU 在内的资源域，其具体操作过程可与图 1 所示方法流程步骤 102 中的操作过程一致。

CPU 调度模块用于根据所述虚拟机的类型，将虚拟机资源域中的 vCPU 调度至物理 CPU 的 CPU 核上执行，其具体操作过程可与图 1 所示方法流程步骤 103 中的操作过程一致。此时，如图 3 所示，CPU 调度模块可具体包括：运行队列加入模块和调度模块。

其中，运行队列加入模块用于根据虚拟机的类型，将虚拟机资源域中的 vCPU 加入到物理 CPU 的 CPU 核的运行队列中。具体实现时，运行队列加入模块可包括集合确定模块和加入模块。其中，集合确定模块用于根据虚拟机的类型，确定虚拟机资源域中的 vCPU 可以加入的 CPU 核的集合；加入模块用于从所述确定的 CPU 核集合中，选取负载最轻的 CPU 核，将 vCPU 加入至所选取的 CPU 核的运行队列中。

调度模块用于在触发虚拟 CPU 调度时，将运行队列中的 vCPU 映射至 CPU 核上执行。

此外，与图 1 所示方法相对应，具体实现时，该多路多核服务器还可如图 3 中的虚线部分所示，进一步包括：虚拟 CPU 势能计算模块和虚拟 CPU 排序模块。

其中，虚拟 CPU 势能计算模块，用于在达到预设时间间隔时，根据预先设置的虚拟机的资源域占用所有物理 CPU 使用率的权重及资源域内的 vCPU 数量，计算运行队列中的 vCPU 的势能。

虚拟 CPU 排序模块用于按照运行队列中 vCPU 的势能，对 vCPU 进行降序排列。

此时，调度模块在触发虚拟 CPU 调度时，根据虚拟机的类型及运行队列头部的虚拟 CPU 势能，将所述运行队列中的虚拟 CPU 映射至 CPU 核上执行。

其中，具体调度过程可参见图 1 所示步骤 103 中描述的调度过程。

本发明实施中，虚拟机监控器内部的各模块可以是物理功能模块，也可

以是软件功能模块，并且各模块还可进行细分或进行合并，具体实现时，本领域普通技术人员可根据实际情况进行处理，此处不再一一列举。

可见，本发明实施例中，根据多路多核服务器上的应用特点，可以创建不同类型的虚拟机，用以满足用户的要求，同时实现计算系统虚拟化效率的最大化。

本发明实施例的主要优点在于：基于半虚拟化原理，利用现有技术的优势，具备良好的提升服务器系统性能的基础；其次，考虑了多路多核服务器中多核处理器的特性，以矩阵表示法标识每个处理器核，可以有效反映 CPU 核间的关系；第三，针对服务器应用的特点，将应用分为吞吐型和并发型两大类应用，相应提出两种不同的 CPU 虚拟化策略，可以同时兼顾虚拟机个体的性能和计算系统整体的效能，达到全面优化系统性能的目的。

基于上述的实现技术，可以实现多路多核服务器的 CPU 虚拟化，考虑了多核处理器的特性和服务器应用的特点，可以获得良好的系统整体效能提升，同时有效提高虚拟机个体的性能。

以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

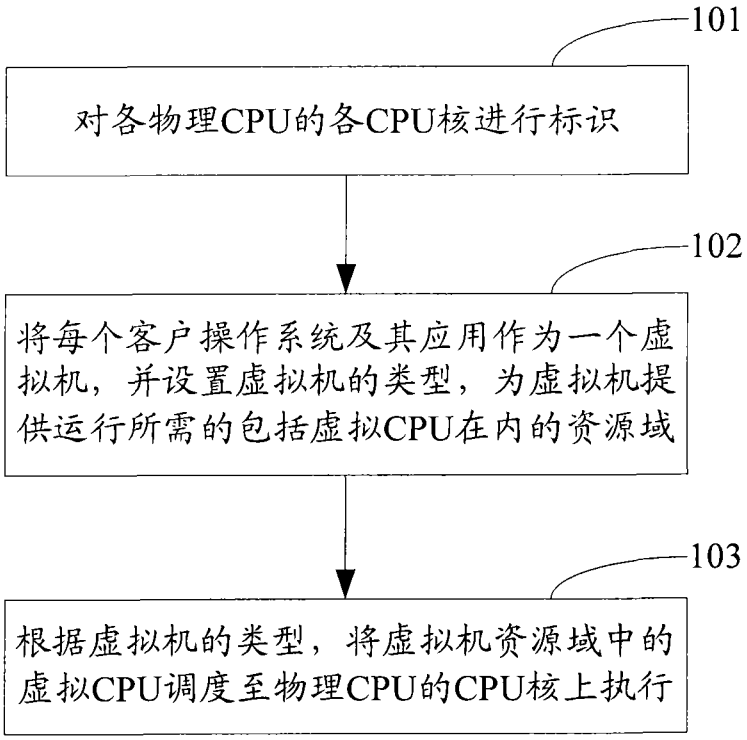

101
对各物理CPU的各CPU核进行标识
102
将每个客户操作系统及其应用作为一个虚拟机，并设置虚拟机的类型，为虚拟机提供运行所需的包括虚拟CPU在内的资源域
103
根据虚拟机的类型，将虚拟机资源域中的虚拟CPU调度至物理CPU的CPU核上执行

图 1

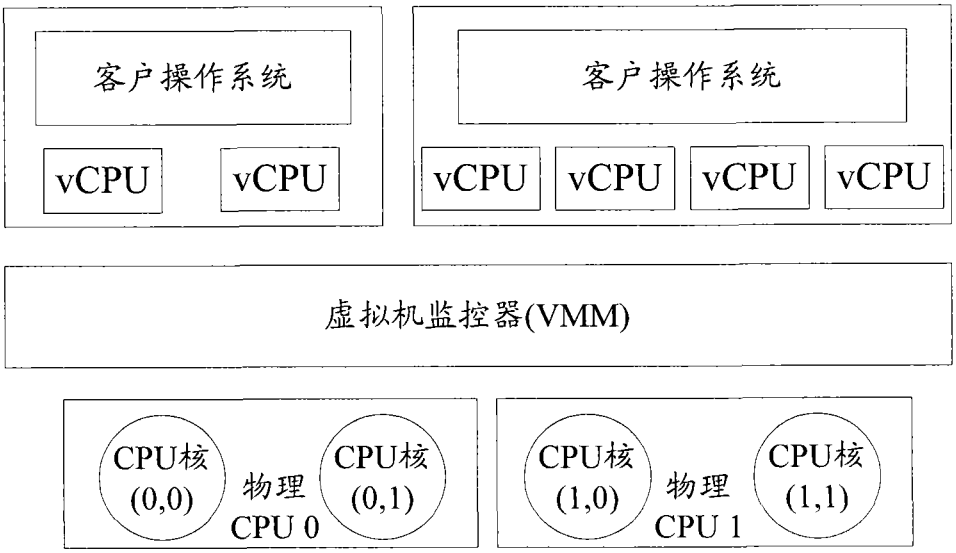

客户操作系统
vCPU
vCPU
客户操作系统
vCPU
vCPU
vCPU
vCPU
虚拟机监控器(VMM)
CPU核 (0,0)
CPU核 (0,1)
物理 CPU 0
CPU核 (1,0)
CPU核 (1,1)
物理 CPU 1

图 2

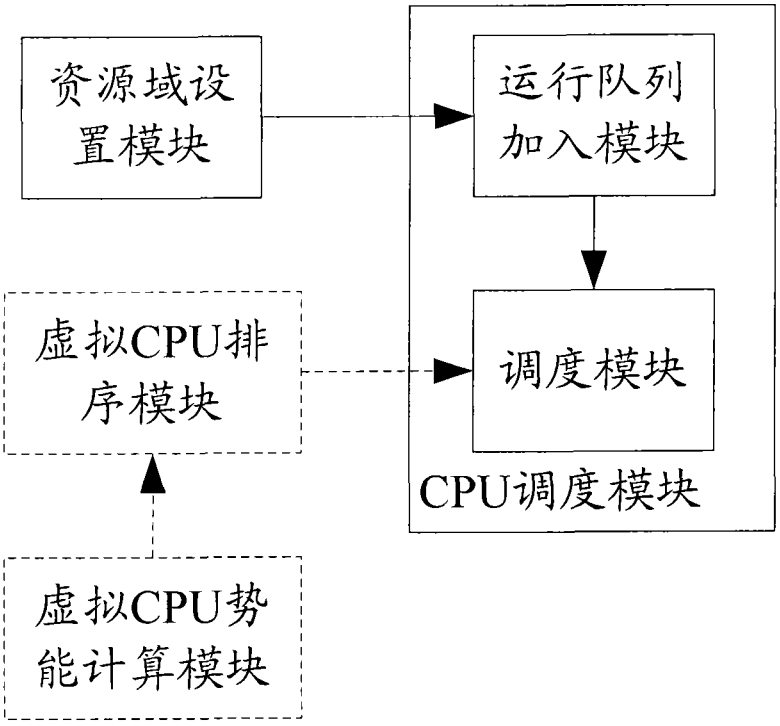


图 3