



(12)发明专利

(10)授权公告号 CN 103597449 B

(45)授权公告日 2017.06.20

(21)申请号 201280028013.9

(22)申请日 2012.06.07

(65)同一申请的已公布的文献号

申请公布号 CN 103597449 A

(43)申请公布日 2014.02.19

(30)优先权数据

13/155,387 2011.06.08 US

(85)PCT国际申请进入国家阶段日

2013.12.06

(86)PCT国际申请的申请数据

PCT/US2012/041434 2012.06.07

(87)PCT国际申请的公布数据

W02012/170746 EN 2012.12.13

(73)专利权人 微软技术许可有限责任公司

地址 美国华盛顿州

(72)发明人 J·阿塞姆

(74)专利代理机构 上海专利商标事务所有限公司 31100

代理人 范玮

(51)Int.Cl.

G06F 9/46(2006.01)

G06F 9/22(2006.01)

(56)对比文件

US 2009037911 A1, 2009.02.05, 全文.

US 2003204780 A1, 2003.10.20, 全文.

US 2005015661 A1, 2005.01.20, 全文.

CN 1755587 A, 2006.04.05, 全文.

审查员 张娟

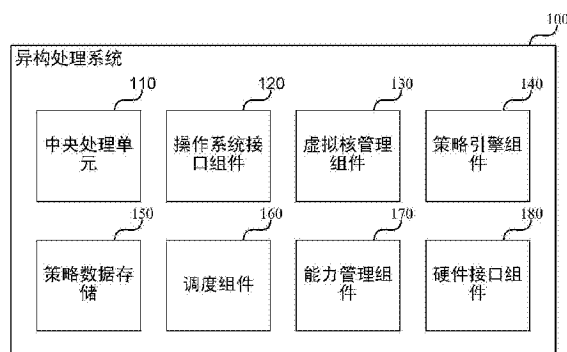
权利要求书2页 说明书8页 附图3页

(54)发明名称

操作系统解耦的异构计算

(57)摘要

本文描述了异构处理系统,该异构处理系统提供软件管理程序以跨大核和小核来自主地控制操作系统线程调度,而无需操作系统知晓或参与,以提高能量效率或满足其他处理目标。该系统向操作系统呈现虚拟化计算核的有限集,其中系统将线程调度到这些虚拟化计算核以供执行。随后,管理程序智能地控制对于哪个或哪些核执行每一线程进行物理分配和选择,以管理能量使用或其他处理要求。通过使用软件管理程序来抽象底层的大和小的计算机体系结构,这些核之间的性能和功率操作差异保持对操作系统不透明。该固有的间接性还将具有新能力的硬件的发行与操作系统发行时间表解耦。



1. 一种用于通过管理异构处理核的管理程序来调度一个或多个操作系统线程的计算机实现的方法,所述方法包括:

由管理程序向操作系统呈现一个或多个虚拟核,其中每个虚拟核将所述操作系统与计算设备可访问的物理处理核之间的一个或多个能力差别相隔离;

从所述操作系统接收对于在管理程序所呈现的所述一个或多个虚拟核中标识出的虚拟核上运行线程的指令的线程调度请求;

确定接收到的调度请求的一个或多个处理需求;

访问指定用于操作所述计算设备的一个或多个目标的调度策略;

选择在其上执行与接收到的调度请求相关联的线程的物理处理核,其中该选择基于所访问的调度策略以及可用系统设施来作出;以及

调度所述线程以在所选择的物理处理核上执行。

2. 如权利要求1所述的方法,其特征在于,接收线程调度请求包括不管操作系统为执行线程而选择的特定虚拟核如何,管理程序根据一个或多个管理程序策略来选择任何特定的物理处理核以执行所述线程。

3. 如权利要求1所述的方法,其特征在于,确定处理需求包括确定特定指令集、可用的协同处理器、指令集扩展、或被调度的线程所使用的体系结构扩展。

4. 如权利要求1所述的方法,其特征在于,确定处理需求包括确定所述线程的一个或多个性能要求。

5. 如权利要求1所述的方法,其特征在于,确定处理需求包括确定所述线程是否适合于以较低功率使用的较慢执行。

6. 如权利要求1所述的方法,其特征在于,确定处理需求包括确定所述线程是否能被延迟,直到附加处理资源可用。

7. 如权利要求1所述的方法,其特征在于,确定处理需求包括访问与特定操作系统有关的专用知识,以确定特定线程的处理需求。

8. 如权利要求1所述的方法,其特征在于,访问调度策略包括确定所述策略请求优化功率使用,并且指定将线程调度到可用的较低功率的处理核。

9. 如权利要求1所述的方法,其特征在于,访问调度策略包括确定所述策略请求优化性能,并且指定将线程调度到可用的较高性能的处理核。

10. 如权利要求1所述的方法,其特征在于,访问调度策略包括确定所述计算设备是具有有限可用功率的移动设备或是具有热限制的设备,并且通过对在其上调度线程的核的选择来指定优化功率使用或管理热。

11. 如权利要求1所述的方法,其特征在于,选择所述物理处理核包括选择系统能在其上调度所述线程的、具有不同能力和性能/功率特性的多个可用核中的一个。

12. 如权利要求1所述的方法,其特征在于,选择所述物理处理核包括以提升用于管理所述计算设备的性能和功率使用的一个或多个目标的方式来作出该选择。

13. 如权利要求1所述的方法,其特征在于,还包括处理所述线程与所选择的物理处理核之间的一个或多个能力差别。

14. 如权利要求1所述的方法,其特征在于,调度所述线程以供执行还包括处理输出并且将所述输出提供回给所述操作系统,从而使所述输出对于所述操作系统而言看上去像是

来自所述操作系统向其分配所述线程的虚拟核。

15. 一种用于通过管理程序来提供操作系统解耦的异构计算的计算机系统,所述系统包括:

一个或多个处理复合体,所述一个或多个处理复合体包括具有异构处理能力和功率概况的一个或多个处理核;

操作系统接口组件,所述操作系统接口组件在管理程序与操作系统之间进行通信,以接收用于向硬件资源递送以及用于接收来自所述硬件资源的输出的指令,以便使不同的核对于所述操作系统而言看上去是一样的;

虚拟核管理组件,所述虚拟核管理组件管理由所述管理程序向所述操作系统呈现的一个或多个虚拟核,其中每个虚拟核将所述操作系统与计算设备可访问的物理处理核之间的一个或多个能力差别相隔离;

策略引擎组件,所述策略引擎组件基于可用的一个或多个处理复合体来管理用于调度操作系统线程以及向所述操作系统呈现虚拟核的一个或多个策略;

调度组件,所述调度组件将作为线程从所述操作系统接收到的一个或多个指令流调度到所述计算设备中所安装的处理复合体中的一个或多个,其中所述处理复合体中的一个或多个基于所访问的调度策略以及可用系统设施被选择;以及

硬件接口组件,所述硬件接口组件在所述管理程序与所述处理复合体之间进行通信,以调度软件指令来在可用物理核上运行。

操作系统解耦的异构计算

[0001] 背景

[0002] 从移动电话到数据中心,能量效率越来越成为重要的区别之处。顾客愿意为更持久的移动设备体验支付额外费用,而且渴望从这些相同设备获取提升的性能。在该规模的另一端,数据中心继续扩大计算能力,然而面临对于什么可被有效地冷却的热限制。另外,公众越来越意识到能量使用和能量使用的环境影响。因此,在许多类型的计算系统中,对能力进行高效使用是较高优先级的设计目标。

[0003] 这些技术上相对的议程(给予更多性能但使用更少功率)导致使用异构设计来进行工业实验,异构设计中在单个系统或硅芯片内,“大”计算核紧密耦合有“小”计算核,在这里被称为异构核或处理。大核(big core)被设计成以较大功率包络提供高性能,而小核(little core)被设计成以较小功率包络提供较低性能。常规做法是操作系统的调度器随后将取决于工作负载来在大核或小核上选择性地调度线程。在一天中至少某些时间期间,操作系统可能能够完全关闭大核并且依赖省电的小核。

[0004] 大核和小核可以或可以不共享相同指令集或特征。例如,小核可包括精简指令集或涉及由操作系统作出进一步决定以在兼容核上调度进程的其他不同。一个传统示例是包括中央处理单元(CPU)和图形处理单元(GPU)且在其空闲或未被充分利用时允许GPU被用于计算任务的系统。

[0005] 现有和当前的解决方案取决于修改操作系统内核,以便“启发”操作系统有大核和小核存在、它们相应的性能和功率特性、以及操作系统可监视系统中的哪些设施(例如,CPU性能计数器、高速缓存未命中/命中计数器、总线活动计数器等)来确定在哪个或哪些核上调度特定线程。这一方法具有若干缺点:1)该方法涉及针对所有被支持的操作系统修改内核,2)该方法需要经修改的内核跨潜在不同的体系结构(例如,支持N个不同实现)来理解大/小设计中的不同,以及3)该方法将操作系统内核的发行时间表与底层计算机体系结构紧密地耦合。于是,计算机体系结构的改变涉及在内核能支持市场上的新核(或反之亦然)之前,等待下一所调度的操作系统发行(即,可能若干年或更久)。

发明内容

[0006] 本文描述了异构处理系统,该异构处理系统提供软件管理程序以跨大核和小核来自主地控制操作系统线程调度,而无需操作系统知晓或参与,以提高能量效率或满足其他处理目标。该系统向操作系统呈现虚拟化计算核的有限集,其中系统将线程调度到这些虚拟化计算核以供执行。随后,在表面下,管理程序智能地控制对于哪个或哪些核(大或小)执行每一线程进行物理分配和选择,以管理能量使用或其他处理要求。通过使用软件管理程序来抽象底层的大和小的计算机体系结构,这些核之间的性能和功率操作差异保持对操作系统不透明。该固有的间接性还将具有新能力的硬件的发行与操作系统发行时间表解耦。硬件供应商可发行更新的管理程序,并且允许新硬件与供应商选择的任何操作系统版本一起工作。

[0007] 提供本发明内容是为了以简化的形式介绍将在以下详细描述中进一步描述的选

择的概念。本发明内容并不旨在标识所要求保护主题的关键特征或必要特征，也不旨在用于限制所要求保护主题的范围。

附图说明

[0008] 图1是示出在一个实施例中的异构处理系统的各组件的框图。

[0009] 图2是示出在一个实施例中的异构处理系统用于使用异构处理核与操作系统之间的管理程序来初始化具有这些核的计算设备的处理的流程图。

[0010] 图3是示出在一个实施例中的异构处理系统用于通过管理异构处理核的管理程序来调度一个或多个处理系统线程的处理的流程图。

[0011] 图4是示出在一个实施例中的异构处理系统的操作环境的框图。

[0012] 详细描述

[0013] 本文描述了异构处理系统，该异构处理系统提供软件管理程序以跨大核和小核来自主地控制操作系统线程调度，而无需操作系统知晓或参与来提高能量效率或满足其他处理目标。该系统向操作系统呈现虚拟化计算核的有限集，其中系统将线程调度到这些虚拟核计算核以供执行。随后，在表面下，管理程序智能地控制对于哪个或哪些核(大或小)执行每一线程进行物理分配和选择，以管理能量使用或其他处理要求。通过使用软件管理程序来抽象底层的大和小的计算机体系结构，这些核之间的性能和功率操作差异保持对操作系统不透明。该固有的间接性还将具有新能力的硬件的发行与操作系统发行时间表解耦。硬件供应商可发行更新的管理程序，并且允许新硬件与供应商选择的任何操作系统版本一起工作。

[0014] 管理程序实现被紧密地耦合到底层计算机体系结构，并且使用可用的系统反馈(例如，CPU利用率、总线/高速缓存活动等)来自主地为所请求的工作负载分配适当的核。这一方法允许底层计算机体系结构与软件管理程序协作来频繁改变，并且将这一进展与上述操作系统解耦。异构处理系统提供简单、粗粒度的功率管理，而无需修改操作系统内核本身。由此，异构处理系统允许更迅速的硬件革新，且允许现有数据中心和其他安装如今从可用的异构处理硬件中受益。

[0015] 异构计算是工业中的新兴领域，其目标是基于系统中可用的不同类型的计算核(例如，CPU、GPU、加速器等)来优化工作负载的执行。优化可以是性能、功率、等待时间或其他目标。尽管适用于这些更一般的情况，但异构处理系统在具有相同功能等效性但不同性能/功率操作特征的核的系统处也是可作为目标的。通常，这些系统具有一个或多个大核以及一个或多个小核。这些大核通常具有深流水线、无序执行、大高速缓存、高时钟速度，并且使用较高泄漏处理(例如，40G)来制造。这些小核通常具有较短的流水线、较小的高速缓存、较低的时钟速度、各种功率级，并且使用低的泄漏处理(例如，40LP)来制造。

[0016] 在某些实施例中，大核和小核可具有体系结构等效性、微体系结构等效性、全局中断控制器、相干性和虚拟化。体系结构等效性可包括相同的指令集体系结构(ISA)、单指令多数据(SIMD)、浮点(FP)、协同处理器(co-processor)可用性、ISA扩展。微体系结构等效性可包括性能上的不同，但相同的可配置特征(例如，高速缓存线长度)。全局中断控制器提供管理、处理中断以及将中断转发到所有核的能力。相干性意味着所有核可存取(高速缓存)来自其他核的数据，并在所需时进行转发。虚拟化是针对切换/迁移来自/去往核的负载。

[0017] 在某些实施例中,异构处理系统可以能够处理核中的微小差异。例如,不支持单指令多数据(SIMD)流扩展(SSE)(现在存在于四个迭代中:SSE1、SSE2、SSE3和SSE4)的小核仍然可处理其他基于Intel x86的软件代码。管理程序可检测指令流中不被支持的指令,并且唤醒向其分配这些流的适当的核。其他指令流可在任何核上忠实地操作。在某些情况下,诸如其中使用了一些不被支持的指令,管理程序可包括某一级的仿真来对可用指令集上不被支持的指令进行仿真。例如,诸如向量数学之类的操作通常可被分解并且使用标准数据指令以较低的效率来实现。

[0018] 在操作系统(OS)初始化之前,软件管理程序在设备引导过程期间安装其自身。在完成指定硬件配置(即,配置存储器、初始化虚拟化设施等)之后,管理程序随后通过策略配置安装在计算设备中的大处理核和小处理器。例如,如果设备是移动电话,则策略可规定管理程序以最少量的可用性能来开始,并且针对电池寿命优化;管理程序随后会将操作系统线程调度到一个或多个小核。或者,如果设备是数据中心刀片,则策略可规定管理程序以大量的可用性能来开始,并且牺牲能量效率;管理程序随后会将操作系统线程调度到可用的大核——以及可能地,小核(取决于可用的热预算)。在实现初始化之后,软件管理程序加载操作系统引导管理器,该操作系统引导管理器随后加载操作系统。

[0019] 在运行时期间,异构处理系统向操作系统呈现虚拟化的内核集。这些核之间的操作特性和区别对操作系统不透明,并且由软件管理程序基于定义的操作策略来私下地管理。操作策略可在系统初始化期间被设置或在运行时期间被动态地设置。

[0020] 管理程序使用操作策略结合可用系统设施(例如,CPU性能计数器、高速缓存未命中/命中计数器、总线活动计数器等)来确定要将操作系统线程调度到哪些核。管理程序将使用该信息来理解CPU核利用率、随时间的趋势、信息位置、以及输入/输出(I/O)模式。根据该信息,管理程序在适当时可动态地且推测性地跨大核和小核迁移操作系统线程。另外,取决于系统实现,管理程序还可代表操作系统来控制动态频率和电压缩放(DVFS)。

[0021] 这里是管理程序可控制的可用操作策略的取样:最小功率(MiPo),最大性能(MaPe),最小功率、按需性能(MiPoD),以及最大性能、空闲时断电(MaPeI)。这些中的每一者在以下段落中描述。然而,附加的、更高级的操作策略可在由任何特定实现选择时被实现。

[0022] 最小功率(MiPo)将线程调度到核的最小集合。通常这将意味着管理程序将线程调度到小核,并且在需要时使用DVFS针对该核来控制功率和性能操作点。可在需要时对附加的小核供电并且调度。

[0023] 最大性能(MaPe)将线程调度到核的最大集合。通常这将意味着管理程序将线程调度到所有的可用核(以大核开始),并且在需要时使用DVFS针对这些核来控制功率和性能操作点。如可用热预算所允许的那样多的小核也被供电并被调度。

[0024] 最小功率、按需性能(MiPoD)通常以最低可用功率状态(例如,在一个或多个小核上)操作,但按工作负载所需来提升性能。这通常被称为操作的“涡轮(turbo)”或“加速(boost)”模式,并且通过动态地分配以及调度到大核来启用。一旦完成工作负载,系统就返回到最小功率状态(例如,在小核上)。

[0025] 最大性能、空闲时断电(MaPeI)通常以最大可用性能状态来操作(例如,在一个或多个大核上),然而一旦达到空闲阈值就默许较低功率状态。在这种情况下,空闲阈值通常不是接近零的CPU利用率,而是可以以某种Dhrystone每秒百万条指令(DMIPS)任意地定义

的,或是由策略定义的CPU利用率百分比。当进入空闲时,管理程序动态地分配并调度到小核,并且将未使用的大核置于待机/停放状态。策略和/或未来的工作负载确定系统何时返回到最大可用性能状态(例如,在大核上)。

[0026] 图1是示出在一个实施例中的异构处理系统的各组件的框图。系统100包括一个或多个中央处理单元110、操作系统接口组件120、虚拟核管理组件130、策略引擎组件140、策略数据存储150、调度组件160、能力管理组件170、以及硬件接口组件180。这些组件中的每个都在此予以更详细讨论。可在位于操作系统与计算设备的硬件资源之间的软件管理程序中实现以下组件。

[0027] 一个或多个中央处理单元110包括具有异构处理能力和功率概况(power profiles)的一个或多个处理核。通常,每一CPU复合体位于单个硅模(silicon die)上,并且CPU复合体的每一核共享一个硅模。硬件可在针对各种类型的设备的各种封装中被实现。例如,出于两者之间的高效通信以及较低的功率使用,较新的移动设备以及甚至某些新近的桌面处理器在同一芯片上包括CPU和GPU。每一CPU复合体可包括一个或多个大核和小核。另选地或另外地,一个CPU复合体可包括所有的大核,而另一CPU复合体包括所有的小核。这里使用的CPU复合体适用于GPU以及可执行软件指令的其他硬件。

[0028] 操作系统接口组件120在管理程序与操作系统之间进行通信,以接收用于向硬件资源递送以及用于接收来自硬件资源的输出的指令。操作系统可调度线程,提供指向指令流的指针(例如,程序计数器(PC)),写入将指令传递到硬件的存储器区域等等。操作系统通常在计算设备上与硬件直接交互。然而,出于各种目的,管理程序在操作系统与硬件之间插入间接层。通常,管理程序被用于提供虚拟化,使得多个操作系统可同时运行在相同硬件上。管理程序还可被用于向操作系统呈现虚拟硬件,虚拟硬件不同于计算设备上所安装的实际硬件。在异构处理系统100的情况下,这可包括使大核和小核对操作系统而言看上去是一样的。系统100甚至可向操作系统呈现与设备上实际存在的相比不同数量的核。

[0029] 虚拟核管理组件130管理由管理程序向操作系统呈现的一个或多个虚拟核。虚拟核对于操作系统而言看上去像是CPU核,但虚拟核与计算设备上可用的物理硬件在特征上可以有所不同。例如,虚拟核可隐藏在处理或功率能力上与操作系统的差异,使得未被设计成与异构大核和小核一起工作的操作系统能够以操作系统被设计的方式来操作。在这些情况下,管理程序提供利用该异构计算环境所需的任何专用编程,使得操作系统无需被修改。

[0030] 策略引擎组件140基于可用的一个或多个中央处理单元来管理用于调度操作系统线程以及向操作系统呈现虚拟核的一个或多个策略。策略引擎组件140可包括专用于特定管理程序实现的硬编码策略,或者可包括可被修改以适合特定安装目标的管理员可配置策略。策略可确定首先调度哪些核,功率使用与处理目标之间的折衷,如何关闭和唤醒核以节省功率、如何向操作系统呈现虚拟核等等。

[0031] 策略数据存储150将一个或多个策略存储在管理程序在引导和执行时可访问的存储设施中。策略数据存储150可包括一个或多个文件、文件系统、硬盘驱动器、数据库、或用于跨系统100的执行会话来持久存储数据的其他存储设施。在某些实施例中,管理员执行设置步骤,该设置步骤通过配置阶段使得系统100存储策略的初始集合供管理程序使用。

[0032] 调度组件160将作为线程从操作系统接收到的一个或多个指令流调度到计算设备上所安装的中央处理单元中的一个或多个。调度组件从操作系统接收标识虚拟核的虚拟核

标识,操作系统请求将线程调度到该虚拟核。调度组件160检查调度请求并且确定在其上调度要执行的线程的物理核。例如,组件160可确定功率或处理对于该线程是否更相关,并且作为响应调度到合适的小核或大核。在某些情况下,组件160可避免将线程调度到特定核来允许那些核被断电以节省功率。

[0033] 能力管理组件170可任选地管理大处理核与小处理核之间的一个或多个差异。在某些情况下,系统100可仅在其中大核和小核共享相同能力的处理单元上操作,而不需要能力管理组件170。在其他情况下,系统100处理可用处理核之间的次要或主要差异。例如,系统100可监视某些核不支持的指令,并且调度的确支持那些指令的核上的对应线程。在更复杂的实现中,组件170可在小核上虚拟化或仿真大核能力(或反之亦然),以满足功率或其他概况的目标。

[0034] 硬件接口组件180在管理程序与中央处理单元之间进行通信,以调度软件指令来在可用物理核上运行。硬件接口组件180可包括对其他组件隐藏且尤其对管理程序所管理的客操作系统隐藏的、真实存储器地址或用于访问真实硬件的其他设施。

[0035] 其上实现异构处理系统的计算设备可包括中央处理单元、存储器、输入设备(例如,键盘和定点设备)、输出设备(例如,显示设备)和存储设备(例如,盘驱动器或其他非易失性存储介质)。存储器和存储设备是可以是编码有实现或启用该系统的计算机可执行指令(如软件)的计算机可读存储介质。此外,数据结构和消息结构可被存储或经由诸如通信链路上的信号等数据传输介质传送。可以使用各种通信链路,诸如因特网、局域网、广域网、点对点拨号连接、蜂窝电话网络等。

[0036] 该系统的实施例可以在各种操作环境中实现,这些操作环境包括个人计算机、服务器计算机、手持式或膝上型设备、多处理器系统、基于微处理器的系统、可编程消费电子产品、数码相机、网络PC、小型计算机、大型计算机、包括上述系统或设备等中任一种的分布式计算环境、机顶盒、片上系统(SOC)等。计算机系统可以是蜂窝电话、个人数字助理、智能电话、个人计算机、可编程消费电子设备、数码相机等。

[0037] 该系统可以在由一个或多个计算机或其他设备执行的诸如程序模块等计算机可执行指令的通用上下文中描述。一般而言,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。通常,程序模块的功能可在各个实施例中按需进行组合或分布。

[0038] 图2是示出在一个实施例中的异构处理系统用于使用异构处理核与操作系统之间的管理程序来初始化具有这些核的计算设备的处理的流程图。

[0039] 在框210开始,系统接收启动请求以初始化计算设备。例如,基本输入/输出系统(BIOS)、可扩展固件接口(EFI)、引导加载器、或其他初始设备软件可加载并调用实现异构计算系统的管理程序。在某些情况下,管理员会在先前已执行安装阶段以便将管理程序安装在计算设备上,尽管系统也可支持网络引导以及其他通常为计算设备提供的非安装场景。

[0040] 在框220继续,系统枚举计算设备的两个或更多个物理处理核。在某些实现中,至少两个核提供不同的性能和功率使用特性。然而,当不存在不对称的情况下,也可使用该系统。例如,使用针对功率管理的软件管理程序在以下场景中仍然适用:其中模具中具有N个物理CPU,但基于诸如环境温度、机壳形状因数、可用功率的成本之类的外在性,仅有K个可

被操作。在引导时,管理程序可使用这一“策略”信息向操作系统报告K个核的虚拟化集合,并且这可根据每一引导循环而变化。在这一场景中,管理程序会为对称的核执行相同的任务。系统可调用BIOS或其他底层以确定计算设备已安装多少处理器以及什么类型的处理器,并且可执行CPUID或其他类似指令以确定与处理器的处理能力有关的信息。在某些实施例中,系统可包括可扩展性接口,通过可扩展性接口能够实现驱动程序或其他管理程序的扩展,并且可扩展性接口可由管理程序制造者或第三方添加以便为管理程序添加对新的处理硬件的支持,而不必更新管理程序本身。

[0041] 在框230继续,系统确定每一所枚举的处理核的能力。这些能力可包括每一核所提供的一个或多个功率概况,每一核所支持的一个或多个指令,每一核的性能特性等等。系统可利用核自身的信息接口(诸如先前提到的CPUID指令)或者由驱动程序或其他扩展提供给管理程序的信息,来确定每一核的能力。系统使用所确定的能力来向每一核分配与该核兼容的线程,并且以与收到策略和处理目标一致的方式来执行调度。

[0042] 在框240继续,系统标识一个或多个操作系统,管理程序将为该一个或多个操作系统管理针对所枚举的物理核的访问和调度。系统可访问硬盘驱动器、闪存驱动器、或计算设备的其他存储,以确定在管理程序被初始化之后要调用哪个操作系统。管理程序可使用与各个操作系统有关的信息来设计,并且可包括可扩展性,使得新的操作系统可被支持,而无需更新管理程序本身。每一操作系统和操作系统版本可具有不同的调度语义或其他细微差别,管理程序处理这些不同的调度语义或其他细微差别以允许操作系统正确地在虚拟化处理资源上执行。在某些情况下,可向管理程序请求允许多个操作系统共享所枚举的物理处理核,并且策略可指示该共享是如何被处理的。

[0043] 在框250继续,系统访问管理程序策略信息,该管理程序策略信息指定用于在所枚举的物理处理核上调度操作系统线程的一个或多个目标。这些目标可包括性能目标,功率使用目标,或用于确定要在哪个或那些核上执行操作系统线程的其他指引。策略可被存储在与计算设备相关联的存储设备中,被硬编码到管理程序实现中等等。管理程序可通过提供给管理员的管理界面来接收对策略的更新。

[0044] 在框260继续,系统创建一个或多个虚拟核以便展示给所标识的操作系统,其中每一虚拟核将操作系统与所确定的、物理处理核之间在能力上的差别相隔离。例如,异构处理系统可将两个或更多个大核和小核作为单个类型的统一虚拟核呈现给操作系统。在从操作系统接收到在虚拟核上执行线程的调度请求后,系统基于所访问的管理程序策略来确定要为该作业选择哪个物理核。管理程序策略可指定管理程序呈现与物理核相比不同数量的虚拟核,诸如当目标是能够使至少某一较高功率需求的核无缝地断电以支持使用较低功率需求的核的情况。或者,系统仍然可使操作系统知晓的核断电,但在操作系统选择使用它们或使用由较低功率的核无法单独满足的一定量的核情况下唤醒这些核。

[0045] 在框270继续,系统调用所标识的操作系统,并且向操作系统呈现所创建的虚拟核,同时将所标识的操作系统与所枚举的物理处理核相隔离。调用操作系统可包括调用操作系统加载器,并且呈现管理程序来代替通常的BIOS或操作系统下面的其他层。如这里描述的,操作系统像它正在物理硬件上直接运行那样操作,但管理程序位于操作系统与物理硬件之间以执行调度逻辑,而无需操作系统知道。在框270之后,这些步骤结束。

[0046] 图3是示出在一个实施例中的异构处理系统用于通过管理异构处理核的管理程序

来调度一个或多个处理系统线程的处理的流程图。

[0047] 在框310开始,系统从操作系统接收对于在管理程序所呈现的标识出的虚拟核上运行线程的指令的线程调度请求,其中虚拟核将操作系统与计算设备可访问的两个或更多个物理处理核之间的一个或多个能力差别相隔离。操作系统通常包括针对每一所检测的处理核的空闲循环,其中操作系统可调度并放置操作系统想在该核上运行的任何指令。操作系统可以对要在特定处理核上运行的多个应用线程进行时间片操作。不管操作系统选择用于执行线程的特定虚拟核如何,管理程序可根据一个或多个管理程序策略来选择任何特定的物理处理核以执行该线程。

[0048] 在框320继续,系统确定接收到的调度请求的处理需求。例如,系统可确定被调度的线程所使用的特定指令集(例如,一个或多个指令集扩展、协同处理器、或其他能力是否被请求),该线程的性能要求,该线程是否适于以较低功率使用的较慢执行,该线程是否可被延迟直到附加处理资源可用等等。系统可使用与特定操作系统或通过策略接收到的指令有关的专用知识来确定特定线程的处理需求。例如,系统可标识与操作系统的内部操作有关的线程、应用线程等,并且根据策略来处理每一者。

[0049] 在框330继续,系统访问指定用于操作该设备的一个或多个目标的调度策略。例如,策略可请求优化功率使用、性能、或这两者的组合。策略可被存储在与该设备相关联的数据存储中,或被硬编码到管理程序的特定实现中。例如,系统可提供偏爱小处理核的管理程序的低功率使用版本,直到执行高性能任务的线程被操作系统调度。在那时,系统可在大核上调度高性能任务,随后在该任务完成之后指引大核休眠。

[0050] 在框340继续,系统选择在其上执行与接收到的调度请求相关联的线程的物理处理核,其中该选择基于所访问的调度策略来作出。系统可具有不同能力和性能/功率特性的多个可用核,系统可在这些核上调度线程。基于系统对核的选取,计算设备将使用更多或更少的功率,并且将用更多或更少的时间完成线程的执行。工作或调度策略是为允许系统以提升用于管理性能、功率或计算设备的其他特性的一个或多个目标的方式来作出选择。移动设备可偏好较低功率使用,而高性能服务器可偏好较高性能。在某些情况下,策略可基于一天的时间(例如,高峰对照非高峰电力成本)或其他考量因素而不同,使得策略随时间或基于特定条件而变化。

[0051] 在框350继续,系统可任选地处理线程与所选择的物理处理核之间的任何能力差别。例如,如果线程包括所选择的核上不可用的指令,则系统可仿真该指令或者用被所选择的核支持的一个或多个等效指令来替换该指令。管理能力差别给系统增加了大量的复杂度,并且管理程序实现者可选择任何特定实现将支持的处理核之间在能力上多大或多小(如果有的话)的差别。

[0052] 在框360继续,系统调度该线程,以在所选择的物理处理核上执行。系统还处理任何输出并且将该输出提供回给操作系统,从而使该输出看上去像是来自操作系统向其分配线程的虚拟核。因此,操作系统保持不知晓管理程序所管理的核的类型和数量,并且在没有管理程序和异构处理核的系统中使用虚拟核集合,如操作系统往常一样。在框360之后,这些步骤结束。

[0053] 图4是示出在一个实施例中的异构处理系统的操作环境的框图。客操作系统410看到由实现这里描述的异构处理系统的软件管理程序430所呈现的一个或多个虚拟处理核

420。软件管理程序430管理异构处理硬件440,异构处理硬件440包括全局中断控制器450、一个或多个大核460、一个或多个小核470。注意到,尽管出于容易例示的目的而仅示出两种核类型(大和小),但系统可以用任何数量的不同的核来操作。例如,某些处理器封装可包括在功率使用和性能方面逐渐下降的若干处理存储。软件管理程序430管理异构处理硬件440,以便将客操作系统410与有效地使用硬件440所需的任何专用知识或处理相隔离。因此,软件管理程序430允许诸如客操作系统410之类的未经修改的传统操作系统利用较新的异构处理硬件440。管理程序430可被修改以跟上硬件440改变,而操作系统410继续照常操作(尽管具有越来越好的功率/性能特性)。

[0054] 在某些实施例中,在线程已被执行之后,异构处理系统将该线程从一个物理处理核迁移到另一个。在某些情况下,在管理程序决定降低功耗、增加性能、或执行其他策略目标时,一个或多个线程可能已在执行。这些核可共享高速缓存存储或其他设施,使得管理程序可将该线程迁移到另一核,而不影响该线程对数据的访问。因此,管理程序可中断该线程的执行,将该线程的指令流移动到不同的物理处理核,并且在目标核上继续执行。

[0055] 在某些实施例中,异构处理系统采用处理器电压和频率修改,以便在选择不同的核之前将功率降低至或将性能增加至一阈值。例如,系统可在大核上开始执行特定线程,随后通过降低该大核的操作电压来缩回该大核的功率使用,并且最后可将该大核的工作迁移到小核。这允许系统逐渐降低功率使用,以管理热封套或满足策略所指定的其他计算目标。

[0056] 在某些实施例中,异构处理系统允许某些处理任务被迁移到云计算设施。系统可呈现任务可被调度到的云计算设施,正如另一处理核那样。对于适当的任务,系统可能能够将任务从计算设备完全卸载,并且之后向客操作系统返回任务的输出。这可允许系统在计算设备上进入较低功率状态,或者将工作从高峰电力成本的数据中心转移到较低电力成本的某一个。

[0057] 在某些实施例中,异构处理系统处理竞争条件并且采用软件锁定范例来管理操作系统期望。在许多情况下,操作系统基于特定线程之间的相互依赖性或缺乏依赖性来调度线程。软件可利用锁、互斥、信号量、或操作系统所提供的其他同步原语来允许软件代码在具有多个同时执行的线程的环境中正确地操作。异构计算系统确保操作系统对线程安全的保证和其他同步被满足,并且可按确保不引入新的竞争条件或其他问题的方式来引入附加锁或确定线程调度。

[0058] 在某些实施例中,异构处理系统包括硬件管理程序。尽管在本文的示例中已使用软件管理程序,但本领域一般技术人员将认识到为实现计算任务而对硬件或软件的选取通常是可被切换以满足性能或其他目标的实现细节。因此,系统可以用硬件管理程序来实现,并且某些处理单元可被制造以便将系统包括在处理单元自身中。

[0059] 从前面的描述中可以看出,可以理解,此处描述的异构处理系统的特定实施例只是为了说明,但是,在不偏离本发明的精神和范围的情况下,可以进行各种修改。因此,本发明只受所附权利要求限制。

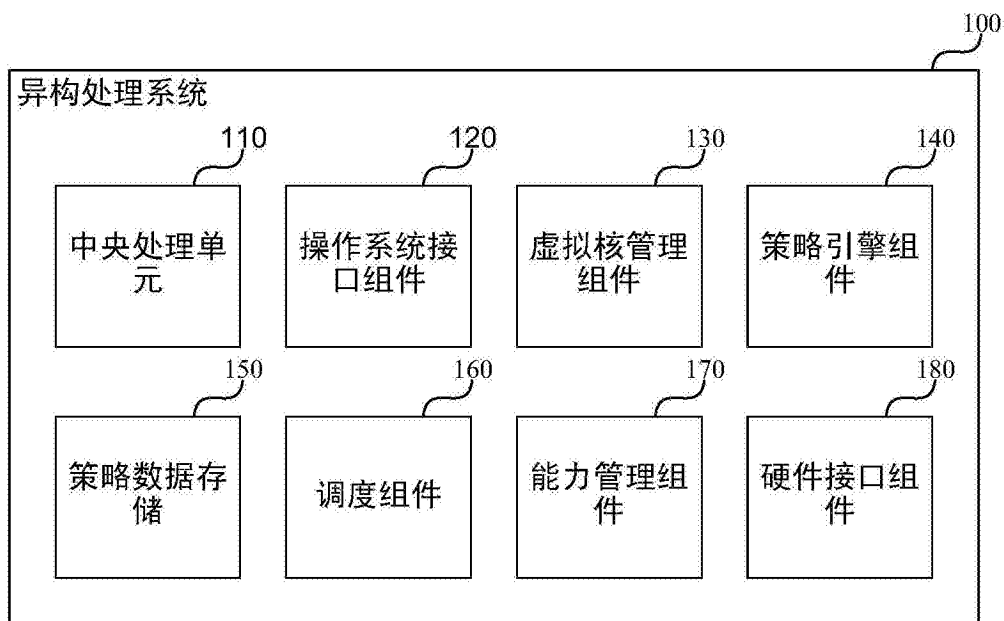


图1

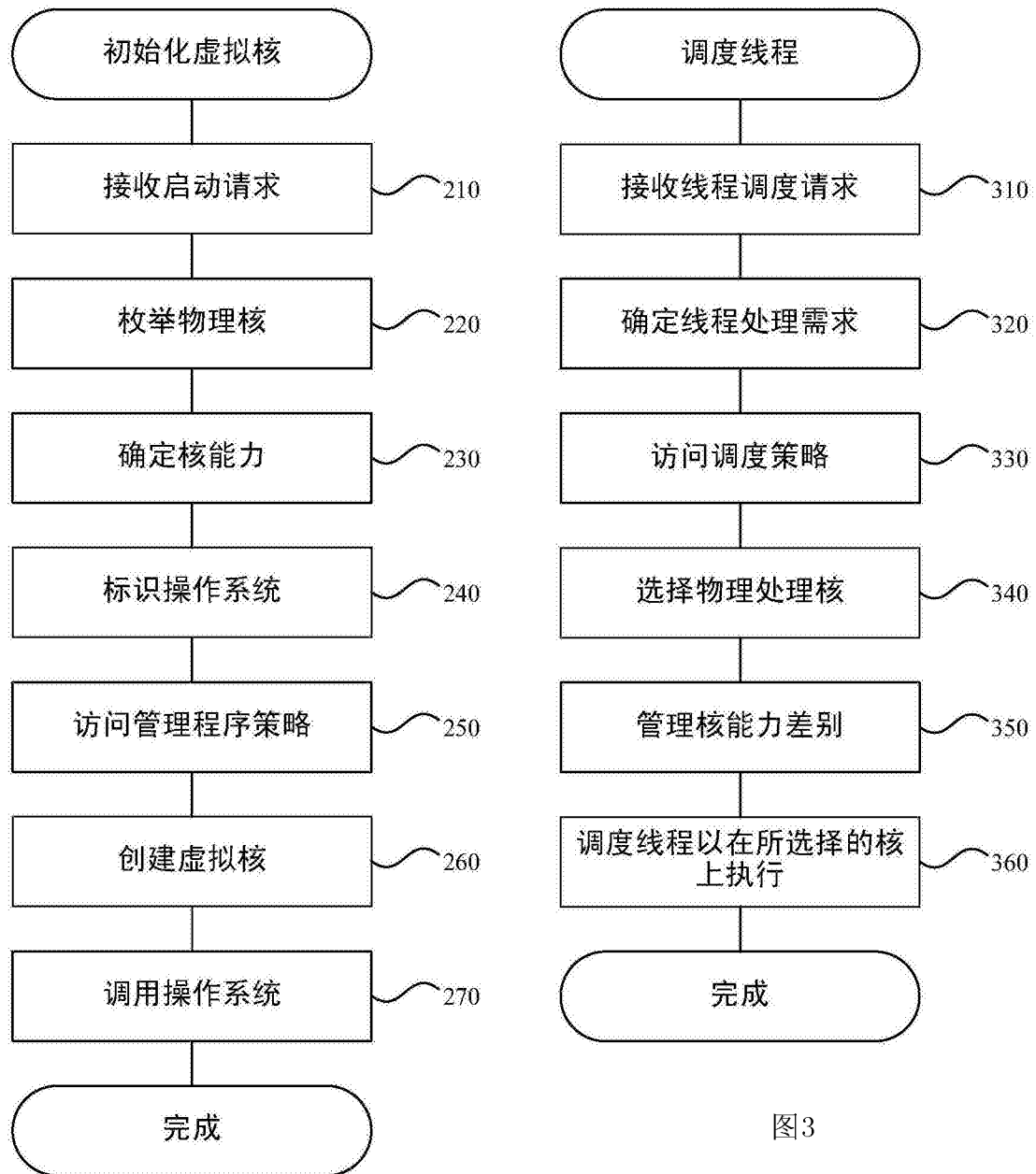


图2

图3

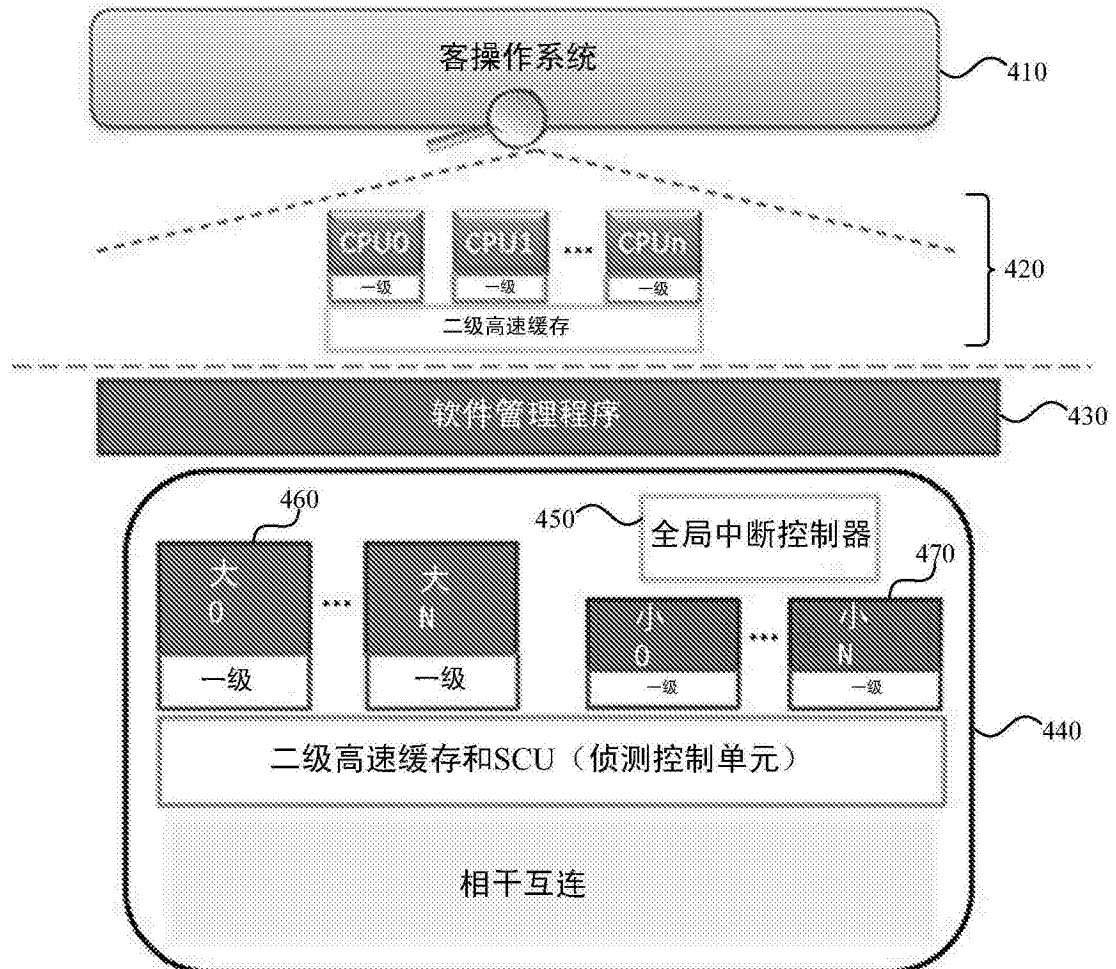


图4