

1. 一种用于在刀片中心中的多刀片机箱之间对工作进行负载平衡的计算机实现的方法,所述计算机实现的方法包括:

经由第一机箱中的管理模块,针对进入所述第一机箱的机箱数据流量监视网络入口交换机;

经由所述第一机箱中的第一刀片上的软件网络代理,针对进入所述第一刀片以及从所述第一刀片排出的第一刀片数据流量监视多个网络接口卡;以及

响应于所述对网络入口交换机的监视以及对第一刀片上的所述多个网络接口卡的监视,通过以下步骤之一确定所述第一机箱被超负荷使用:

(1) 测量所述第一机箱的第一资源使用水平,其中资源使用水平是基于与总共有多少来自于所述第一机箱中的刀片的硬件资源可用相比总共有多少来自于所述第一机箱中的刀片的硬件资源正在被使用;测量第二机箱的第二资源使用水平,其中所述第二资源使用水平小于所述第一资源使用水平;计算所述第一资源使用水平与所述第二资源使用水平之间的差,来确定资源使用水平差;以及如果所述资源使用水平差超过预定义的动作水平,则确定所述第一机箱超负荷使用;

(2) 测量第一刀片机箱数据流量,其中所述第一刀片机箱数据流量是寻址到所述第一刀片并且通过网络入口交换机的数据流量;测量第一刀片网络接口卡数据流量,其中所述第一刀片网络接口卡数据流量是已经由所述第一刀片处理的并且从所述第一刀片排出的数据流量;以及计算所述第一刀片机箱数据流量和所述第一刀片网络接口卡数据流量之间的差,其中所述第一刀片机箱数据流量和所述第一刀片网络接口卡数据流量之间的差超过了预定义的动作水平指示所述第一机箱中的所述第一刀片中资源的过度使用;

从所述第一机箱停用所述第一刀片,以及

启用第二机箱中的第二刀片。

2. 根据权利要求1所述的计算机实现的方法,进一步包括:

将寻址到所述第一刀片的任何未来的数据流量重路由到所述第二刀片。

3. 根据权利要求1所述的计算机实现的方法,其中通过确定网络接口卡上的数据流量超过预定的网络接口卡数据流量水平,而将所述第一刀片识别为将被解耦合的超负荷使用的刀片。

4. 根据权利要求3所述的计算机实现的方法,进一步包括:

响应于所述第一刀片被从所述第一机箱解耦合,利用在所述第一刀片被解耦合之前由所述第一刀片使用的同一操作系统来引导所述第二刀片。

5. 根据权利要求1所述的计算机实现的方法,其中所述第一刀片经由软件机箱接口耦合至所述第一机箱中的第一干线,并且其中所述第一刀片从所述第一机箱的停用是通过禁用所述软件机箱接口的软件来实现的。

6. 根据权利要求1所述的计算机实现的方法,其中所述第一刀片经由硬件插口耦合至所述第一机箱中的第一干线,并且其中所述第一刀片从所述第一机箱的停用是通过由将所述第一刀片从所述第一机箱中的所述第一干线上物理地拔出的线性硬件机构执行的。

7. 根据权利要求6所述的计算机实现的方法,其中所述线性硬件机构由响应于所述第一机箱正被超负荷使用的软件来控制,其中所述线性硬件机构使得被拔出的第一刀片超过所述第一机箱中的其他刀片而从所述第一机箱突出离开,以及其中突出的刀片提供了所述

刀片机箱中的哪个所述刀片已经被物理地拔出的可视指示符。

8. 一种用于在刀片中心中的多刀片机箱之间对工作进行负载平衡的计算机实现的系统,所述计算机实现的系统包括:

用于经由第一机箱中的管理模块,针对进入所述第一机箱的机箱数据流量监视网络入口交换机的装置;

用于经由所述第一机箱中的第一刀片上的软件网络代理,针对进入所述第一刀片以及从所述第一刀片排出的第一刀片数据流量监视多个网络接口卡的装置;以及

用于响应于所述对网络入口交换机的监视以及对第一刀片上的所述多个网络接口卡的监视,通过以下装置之一确定所述第一机箱被超负荷使用的装置:

(1) 用于测量所述第一机箱的第一资源使用水平的装置,其中资源使用水平是基于与总共有多少来自于所述第一机箱中的刀片的硬件资源可用相比总共有多少来自于所述第一机箱中的刀片的硬件资源正在被使用;用于测量第二机箱的第二资源使用水平的装置,其中所述第二资源使用水平小于所述第一资源使用水平;用于计算所述第一资源使用水平与所述第二资源使用水平之间的差,来确定资源使用水平差的装置;以及用于如果所述资源使用水平差超过预定义的动作水平,则确定所述第一机箱超负荷使用的装置;

(2) 用于测量第一刀片机箱数据流量的装置,其中所述第一刀片机箱数据流量是寻址到所述第一刀片并且通过网络入口交换机的数据流量;用于测量第一刀片网络接口卡数据流量的装置,其中所述第一刀片网络接口卡数据流量是已经由所述第一刀片处理的并且从所述第一刀片排出的数据流量;以及用于计算所述第一刀片机箱数据流量和所述第一刀片网络接口卡数据流量之间的差的装置,其中所述第一刀片机箱数据流量和所述第一刀片网络接口卡数据流量之间的差超过了预定义的动作水平指示所述第一机箱中的所述第一刀片中资源的过度使用;

用于从所述第一机箱停用所述第一刀片的装置,以及

用于启用第二机箱中的第二刀片的装置。

9. 根据权利要求 8 所述的计算机实现的系统,进一步包括:

用于将寻址到所述第一刀片的任何未来的数据流量重路由到所述第二刀片的装置。

10. 根据权利要求 8 所述的计算机实现的系统,其中还包括:用于通过确定网络接口卡上的数据流量超过预定的网络接口卡数据流量水平,而将所述第一刀片识别为将被解耦合的超负荷使用的刀片的装置。

11. 根据权利要求 10 所述的计算机实现的系统,进一步包括:

用于响应于所述第一刀片被从所述第一机箱解耦合,利用在所述第一刀片被解耦合之前由所述第一刀片使用的同一操作系统来引导所述第二刀片的装置。

12. 根据权利要求 8 所述的计算机实现的系统,其中所述第一刀片经由软件机箱接口耦合至所述第一机箱中的第一干线,并且其中所述第一刀片从所述第一机箱的停用是通过禁用所述软件机箱接口的软件来实现的。

13. 根据权利要求 8 所述的计算机实现的系统,其中所述第一刀片经由硬件插口耦合至所述第一机箱中的第一干线,并且其中所述第一刀片从所述第一机箱的停用是通过由将所述第一刀片从所述第一机箱中的所述第一干线上物理地拔出的线性硬件机构执行的。

14. 根据权利要求 13 所述的计算机实现的系统,其中所述线性硬件机构由响应于所述

第一机箱正被超负荷使用的软件来控制,其中所述线性硬件机构使得被拔出的第一刀片超过所述第一机箱中的其他刀片而从所述第一机箱突出离开,以及其中突出的刀片提供了所述刀片机箱中的哪个所述刀片已经被物理地拔出的可视指示符。

用于对工作进行负载平衡的方法和系统

技术领域

[0001] 本公开内容涉及计算机领域,并且具体涉及刀片服务器。更具体地,本公开内容涉及在刀片中心中的刀片机箱之间的动态负载平衡。

背景技术

[0002] 虽然早期的计算机架构利用独立的单个计算机(通常称为个人计算机(PC)),但更强大的现代计算机系统经常使用在公共机箱中耦合在一起的多个计算机。一种示例性的公共机箱称为刀片机箱,其包括通过刀片机箱内的公共干线加以耦合的多个刀片。在刀片机箱充当企业的刀片服务器时被称作服务器刀片的每个刀片是可插拔的板,其包括至少一个处理器、板载存储器以及输入/输出(I/O)接口。多个刀片能够彼此通信并共享公共资源,诸如存储设备、监视器、输入设备(键盘、鼠标)等。

[0003] 多个刀片机箱组成刀片中心。刀片中心通常专用于单个企业和/或特定功能,诸如处理贷款、管理工资表等。由于刀片中心包括多个刀片机箱,所以在不同刀片机箱之间经常发生工作负载的不平衡。当前的负载平衡通常仅仅是通过监视去往不同机箱的总数据流量,来将工作从超载的机箱移动到不太繁忙的机箱。此类负载平衡对检测和修正负载不平衡的潜在原因帮助不大。

发明内容

[0004] 为了解决上述问题,提供了一种用于在刀片中心内的不同刀片机箱之间进行负载平衡的计算机实现的方法、系统和计算机程序产品。针对去往和来自刀片机箱的总数据流量监视每个刀片机箱。如果刀片机箱集群内的第一刀片机箱负担过重,则释放第一刀片机箱内负担最重的刀片。继而识别该刀片机箱集群内的较少利用的第二刀片机箱,并且指派该第二刀片机箱内的负担最轻的刀片来接管从第一刀片机箱释放的刀片的工作负载。

[0005] 在以下的详细描述中,本发明的上述以及附加用途、特征和优势将变得明显。

附图说明

[0006] 在所附权利要求书中记载了相信是本发明具有新颖特征的特性。然而,在结合附图阅读时,通过参考对示范实施方式的以下详细描述,将获得对本发明本身以及其优选的使用模式、其他目的和优势的最佳的理解。

[0007] 图1示出了可以由本发明利用的示例性服务器刀片;

[0008] 图2示出了在刀片中心内的刀片机箱之间自动进行负载平衡工作的独特且新颖的刀片中心;

[0009] 图3示出了图2所示的刀片机箱的附加细节;

[0010] 图4是用于在刀片中心内的刀片机箱之间进行负载平衡的示例性步骤的流程图;

[0011] 图5示出了用于物理地以及机械地从刀片机箱解耦合服务器刀片的线性机构;以及

[0012] 图 6 示出了用于智能地以及选择性地将服务器刀片与同一硬盘驱动器耦合 / 解耦合的机械耦合器。

具体实施方式

[0013] 现在参考附图,并且特别地参考图 1,其示出了本发明可以使用的示例性刀片 102 的框图。刀片 102 是服务器刀片机箱(在下面关于图 2 至图 3 的描述中,示出为刀片机箱 202、204 和 / 或 302) 的组件。刀片机箱 202、204 和 / 或 302 通常用作服务器;因此,刀片 102 可以称为服务器刀片。应该指出,针对刀片 102 所示出的一些或所有示例性架构可以由软件部署服务器 150 和 / 或其他刀片 124 使用,其他刀片 124 可以与刀片 102 处于同一刀片机箱内。

[0014] 刀片 102 包括耦合至系统总线 106 的处理器单元 104。处理器单元 104 可以利用每个都具有一个或多个处理器核心的一个或多个处理器。驱动 / 支持显示器 110 的视频适配器 108 也经由到机箱干线 157 的机箱接口 131 耦合至系统总线 106(以下更详细地进行描述)。在一个实施方式中,开关 107 将视频适配器 108 耦合至系统总线 106。备选地,开关 107 可以将视频适配器 108 耦合至显示器 110。在另一实施方式中,开关 107 是开关,优选地为机械开关,其允许显示器 110 耦合至系统总线 106,并且因此仅在执行这里所述方法的指令执行时发挥功能(例如,下面描述的刀片管理软件 BMS 148)。该切换引起刀片 102 从一个其中不显示此处所述方法的未决步骤和 / 或结果的系统实质转变到一个其中显示这些结果的系统。

[0015] 系统总线 106 经由总线桥 112 耦合至输入 / 输出 (I/O) 总线 114。I/O 接口 116 耦合至 I/O 总线 114。I/O 接口 116 担负与各种 I/O 设备的直接通信或经由机箱接口 131 的通信,机箱接口 131 是允许刀片 102 耦合至刀片机箱中的机箱干线 157 的硬件和 / 或软件。一旦耦合至机箱干线 157,则刀片 102 能够与除了显示器 110 之外的其他设备通信,包括键盘 118、鼠标 120、媒体托盘 122(其可以包括诸如 CD-ROM 驱动器、多媒体接口等的存储设备)、处于刀片机箱内的其他刀片 124 以及(如果没有以下面描述的方式使用 VHDL 芯片 137)USB 端口 126。应该指出,虽然示出其他刀片 124 经由机箱接口 131 耦合至刀片 102,但是在一个实施方式中,这些其他刀片 124 可以经由网络 128(特别是如果网络 128 是刀片中心内的局域网 (LAN)) 耦合至刀片 102。虽然连接至 I/O 接口 116 的端口形式可能对与计算机架构领域的技术人员来说是已知的,但是在优选实施方式中,一些或所有这些端口是通用串行总线 (USB) 端口。

[0016] 如上所述,刀片 102 能够与软件部署服务器 150 通信,并且在一个实施方式中,能够使用诸如网络接口卡 (NIC) 130 之类的网络接口经由网络 128 与刀片机箱内的其他刀片 124 通信,网络接口卡 (NIC) 130 直接耦合至系统总线 106,或者备选地经由机箱接口 131 耦合至系统总线 106。网络 128 可以是诸如因特网的外部网络,或者是诸如以太网或虚拟专用网 (VPN) 之类的内部网络。

[0017] 硬盘驱动器接口 132 也耦合至系统总线 106。硬盘驱动器接口 132 与硬盘驱动器 134 接口连接。在优选实施方式中,硬盘驱动器 134 构成系统存储器 136,其也耦合至系统总线 106。系统存储器被定义为刀片 102 中的易失性存储器的最低级别。该易失性存储器包括附加的更高级别的易失性存储器(未示出),包括但不限于高速缓冲存储器、寄存器和

缓冲器。构成系统存储器 136 的数据包括刀片 102 的操作系统 (OS) 138 和应用程序 144。

[0018] OS 138 包括壳 140, 用于提供到诸如应用程序 144 的资源的透明用户访问。通常, 壳 140 是在用户和操作系统之间提供解译器和接口的程序。更具体地, 壳 140 执行被输入到命令行用户接口或来自于文件的命令。因此, 壳 140 还称为命令处理器, 其通常是操作系统软件层次的最高级并且充当命令解译器。壳提供系统提示、解译由键盘、鼠标或其他用户输入介质输入的命令, 以及向操作系统 (例如, 内核 142) 的适当的较低级发送已解译命令用于处理。应该指出, 虽然壳 140 是基于文本的、面向行的用户接口, 但是本发明将同样良好地支持其他用户接口模式, 诸如图形、语音、手势等。

[0019] 如上所述, OS 138 还包括内核 142, 其包括 OS 138 的较低级的功能, 包括由 OS 138 的其他部分和应用程序 144 所需的必要范围, 包括存储器管理、进程和任务管理、磁盘管理以及鼠标和键盘管理。

[0020] 应用程序 144 包括绘制器, 以示例性方式示出为浏览器 146。浏览器 146 包括使用超文本传输协议 (HTTP) 消息发送而支持万维网 (WWW) 客户端 (即, 刀片 102) 发送和接收网络消息到因特网, 因此支持与软件部署服务器 150 和其他所述计算机系统通信的程序模块和指令。

[0021] 刀片 102 的系统存储器 (以及软件部署服务器 150 的系统存储器) 中的应用程序 144 还包括刀片管理软件 (BMS) 148。BMS 148 包括用于实现以下所述处理的代码, 包括在图 2 至图 3 中描述的那些。在一个实施方式中, 刀片 102 能够从软件部署服务器 150 下载 BMS 148, 包括按需下载。还应该进一步指出, 在本发明的一个实施方式中, 软件部署服务器 150 执行与本发明相关联的所有功能 (包括执行 BMS 148), 因此将刀片 102 从必须使用其自己的内部计算资源来执行 BMS 148 中解脱。

[0022] 在系统存储器 136 中还存储了 VHDL (VHSIC 硬件描述语言) 程序 139。VHDL 是一种示例性设计输入语言, 其用于现场可编程门阵列 (FPGA)、专用集成电路 (ASIC) 和其他类似的电子设备。在一个实施方式中, 执行来自于 BMS 148 的指令使得 VHDL 程序 139 配置 VHDL 芯片 137, 该芯片可以是 FPGA、ASIC 等。对 VHDL 芯片 137 进行编程引起刀片 102 架构的实质转变, 其中 (假设 USB 端口 126 没有耦合至 I/O 接口 116, 并且 NIC 130 没有直接耦合至系统总线 106) USB 端口 126 和 NIC 130 现在经由 VHDL 芯片 137 选择性地耦合至系统总线 106。

[0023] 在本发明的另一实施方式中, 执行来自于 BMS 148 的指令导致使用 VHDL 程序 139 来对 VHDL 仿真芯片 151 进行编程。VHDL 仿真芯片 151 可以合并与 VHDL 芯片 137 的上述架构类似的架构。一旦 BMS 148 和 VHDL 程序 139 对 VHDL 仿真芯片 151 编程, 则 VHDL 仿真芯片 151 作为硬件执行由对在 BMS 148 找到的一些或所有指令的一个或多个进行描述的一些或所有功能。即, VHDL 仿真芯片 151 是对在 BMS 148 中找到的一些或所有软件指令的硬件仿真。在一个实施方式中, VHDL 仿真芯片 151 是可编程只读存储器 (PROM), 一旦根据来自于 BMS 148 和 VHDL 程序 139 对其进行烧制, 该 VHDL 仿真芯片 151 被永久转变为执行下面在图 2 中示出的元件 206、208、210、212 和 / 或 214a-b 的功能的新电路。因此, VHDL 仿真芯片 151 还被适当地视为在刀片 102 控制之下的机器。应该指出, 虽然 VHDL 仿真芯片 151 被描述为是与刀片 102 分离的不同实体, 但是在另一实施方式中, VHDL 仿真芯片 151 可以是刀片 102 的整体部分。

[0024] 在刀片 102 中描述的硬件元件不旨在穷举目的,而是本发明所需的必备组件的说明。例如,刀片 102 可以包括备选的存储器存储设备,诸如磁带、数字多功能盘 (DVD)、Bernoulli 盒式磁带等。这些和其他变形旨在处于本发明的精神和范围之内。

[0025] 现在参考图 2,示出了构思的由本发明使用的以及本发明一起使用的示例性刀片中心 200。刀片中心 200 包括至少两个刀片机箱,示出为刀片机箱 202 和刀片机箱 204。在刀片机箱 202 内有多片刀片 214a-n,其中“n”是任何整数,但优选为 14。在刀片机箱 204 内有多片刀片 216a-n,其中“n”是任何整数,并且其中“n”可以大于、小于或等于刀片机箱 202 中的刀片 214 的数量。刀片 214a-n(在不指定为 214 的情况下,每个独立引用为集合/子集)耦合至机箱干线 218,其在刀片 214a-n 中提供机械和逻辑连接(例如,数据和控制信号交换)。类似地,刀片 216a-n 耦合至机箱干线 220。

[0026] 刀片 214a-n 和刀片 216a-n 分别由管理模块 (MM) 210 和 212 管理。每个 MM 包括处理器(未示出),用于控制特定刀片的输入/输出 (I/O) 功能、将特定刀片与网络(例如,网络 222,在图 1 中示出为 128) 接口连接,并且向特定刀片分配任务和数据。

[0027] 每个刀片包含主管代理(分别示出为示例性刀片 214n 的主管代理 215a 和刀片 216a 中的主管代理 215b)。主管代理 215a-b 是与网络使用监视器 208 和/或负载平衡管理器 206 通信的相同软件的副本。网络使用监视器 208 是监视机箱和/或刀片的使用水平(例如,它们的繁忙程度)的软件逻辑。负载平衡管理器 206 通过经由诸如网络入口交换机(示出为网络入口交换机 224 和 226) 的软件交换机而使此类流量寻址到特定刀片,来将来自于网络 222 的工作和数据分组分配给特定刀片。

[0028] 如图 2 所示,主管代理 215(示出为 215a 或 215b) 负责监视特定刀片的活动水平(例如,正在使用的诸如总线带宽、CPU 使用率等内部硬件资源的百分比)。网络入口交换机(示出为 224 或 226)(经由 MM) 向负载平衡管理器 206 和/或网络使用监视器 208 提供流量监视信息,该信息反映全部刀片机箱(例如,特定刀片机箱内的所有刀片的活动水平)的活动水平。

[0029] 如下面将进一步描述的,如果第一机箱中的刀片被认为超负荷使用(负担过重),那么第二机箱中的另一刀片将接管第一机箱中的负担过重的刀片的操作。在优选实施方式中,通过编程地将引导驱动从第一刀片重映射到第二刀片,来停用第一机箱中刀片。例如,假设刀片 214n 是负担过重的刀片。为了释放刀片 214n,禁用引导驱动 228,因此卸下刀片 214n 的操作系统。假设刀片 216a 将接管刀片 214n 的操作,现在指示刀片 216a 来映射到引导驱动 228(存储在引导存储设备 230 中),从而使用与之前由刀片 214n 使用的引导驱动相同的引导驱动 228 来接合 216a。

[0030] 现在参考图 3,示出了刀片(在图 2 中示出的示例性刀片 214n) 的附加细节。如图所示,刀片 214n 使用多个 NIC 302a-n 来与网络 304(在图 1 中示出为网络 128 在图 2 中示出为网络 222) 通信。网络 304 提供去往和来自刀片 214n 的网络流量(例如,数据、指令、任务等)。该流量可以来自于刀片机箱 202 内的其他刀片 214,但是在本发明的环境中优选是来自于刀片机箱 202 的外部,诸如来自于远程工作服务器 306 等。主管代理 215a 监视通过 NIC 302a-n 的流量活动水平。该监视允许主管代理 215a 确定刀片 214n 的繁忙程度。在本发明的一个实施方式中,主管代理 215a 将其监视集中于出站(排出)数据。因此,通过将有多少数据进入刀片 214n(来自网络 304 经由负载平衡管理器 206 和网络入口交换机 224)

与有多少数据流量离开刀片 214n (其中排出数据是刀片 214n 对从网络入口交换机 224 接收分组进行数据处理的结果) 进行比较,可以由负载平衡管理器 206 和 / 或网络使用监视器 208 做出关于刀片 214n 超载的程度的确定。即,如果在任务进入刀片 214n 时和当来自该任务的输出离开刀片 214n 时之间存在过长的延迟 (基于预定的时间长度),那么可以适当地假设刀片 214n 内的资源负担过度,并且因此刀片 214n 是被设置离线以及由另一机箱中的另一刀片 (例如,更强大或不太忙) 替换的良好候选。应该指出,在优选实施方式中,刀片机箱内的一个刀片不替换该同一刀片机箱中的另一刀片,因为优选实施方式的主要目的在于平衡刀片机箱之间的工作量而不仅仅是平衡独立刀片之间的工作量。

[0031] 现在参考图 4,示出了本发明采取的在不同刀片机箱之间或不同刀片机箱中进行负载平衡的示例性步骤的流程图。在初始块 402 之后,在第一机箱中引导第一刀片 (块 404)。使用可以选定的操作系统来引导该第一刀片,该选定的操作系统可以专用于将由管理模块发送到第一刀片的特定类型任务。该第一刀片经由网络入口交换机和多个 NIC 两者耦合至网络 (块 406),如图 3 所示。然后,管理模块针对寻址到第一机箱中的第一刀片的网络流量监视网络入口交换机 (块 408)。此外,第一刀片上的网络代理监视通过第一刀片上 NIC 的网络流量 (块 410)。

[0032] 如查询块 412 所示,确定第一机箱是否超负荷使用。存在可以确定第一机箱是否超负荷使用的不同计算机实现的方法。在一个实施方式中,通过测量第一机箱的第一资源使用水平来做出该确定,其中,资源使用水平是基于与总共有多少来自于所述第一机箱中的刀片的硬件资源可用相比总共有多少来自于所述第一机箱中的刀片的硬件资源正在被使用。测量第二机箱的第二资源使用水平,其中第二资源使用水平小于第一资源使用水平。计算第一资源使用水平与第二资源使用水平之间的差,以确定资源使用水平差。因此,如果资源使用水平差超过了预定义的动作水平,则确定第一机箱超负荷使用。

[0033] 在另一实施方式中,通过首先测量对第一刀片机箱数据流量,来实现确定第一机箱已经超负荷使用,其中,第一刀片机箱数据流量是寻址到第一刀片并且通过网络入口交换机的数据流量。下一步骤是测量第一刀片 NIC 数据流量,其中第一刀片 NIC 数据流量是已经由第一刀片处理并且从第一刀片排出的数据流量。最后,计算第一刀片机箱数据流量与第一刀片 NIC 数据流量之间的差,其中预定大的差指示第一机箱中的第一刀片中的资源过度使用。

[0034] 在一个实施方式中,通过确定 NIC 上的数据流量超过预定的 NIC 数据流量水平,来将第一刀片识别为将被解耦合的超负荷使用的刀片。该预定的 NIC 数据流量水平可以由管理模块、网络使用管理器、网络使用监视器或具有适当监管权限的任何管理员设置。

[0035] 如果第一机箱超负荷使用并且已经识别了负担过重的刀片,则针对未充分使用的第二机箱进行查询 (查询块 414)。使用上述相同的标准和设备进行该查询,以便确定第一机箱超负荷使用。即,如果第二机箱具有去往其的“轻”量网络流量 (如由监视上述过程的网络入口交换机确定),那么确定第二机箱为用于从第一机箱中的负担过重的第一刀片接受工作的美好候选。未来工作 (其本该被发送到第一机箱中的负担过重的第一刀片) 的目的地将是第二机箱中的最少使用的刀片。使用用于识别第一机箱中的负担过重的第一刀片的同一过程来识别该最少使用的刀片。即,将第二机箱中的如下刀片分类为用于接管第一机箱中负担过重的第一刀片的工作负担的美好候选,其中该刀片来自于第二机箱中的网络

入口交换机的网络流量为轻量,并来自于该刀片的 NIC 流量为轻量。

[0036] 一旦已经识别了第一机箱中的负担过重的第一刀片以及第二机箱中的替换该第一刀片的第二刀片,则允许第一刀片完成任何正在运行的任务(块 416)。这避免与上下文切换相关联的开销,而这将引起未决工作负担的意外取消。然而,一旦完成了未决任务(查询块 418),则从第一机箱释放第一刀片(块 420)。

[0037] 如下所述,一旦使第一刀片释放(停用),则另一机箱中的第二刀片将接合(启用)。在本发明的优选实施方式中,通过编程地将引导驱动从第一刀片重映射到第二刀片来执行该停用/启用,如图 2 所示。如上所述,通过卸下负担过重的第一刀片的引导程序(以及操作系统)来停用该负担过重的第一刀片,从而禁用该负担过重的第一刀片。同时,将第二机箱中的失效备援(failover)刀片映射到同一引导驱动,其存储在外部存储系统中(例如,图 2 所示的引导存储 230)。该同一引导驱动(图 2 中所示的引导驱动 228)是在第一机箱中的负担过重的第一刀片在被卸下之前使用的同一引导驱动。

[0038] 备选地,可以单独地或组合地使用四个备选过程,来释放/停用/禁用第一刀片并且接合/启用/使能失效备援刀片。第一释放方法是经由指示第一机箱中的管理模块不向第一刀片发送更多工作的软件。第二方式是经由软件机箱接口到第一机箱中的第一干线,使得通过禁用该软件机箱接口的软件实现将第一刀片从第一机箱的解耦合。第三方法使用耦合开关(在图中未示出),其是将刀片耦合至机箱干线/将刀片从机箱干线解耦合的软件控制的硬件开关。第四方法使用线性运动硬件,其中第一刀片经由硬件插口耦合至第一机箱中的机箱干线,并且其中通过物理上将第一刀片从第一机箱中的第一干线拔出的线性硬件机构,来执行将第一刀片从第一机箱的解耦合。如果使用第四方法,可以由响应于第一机箱超负荷使用的软件控制该线性硬件机构。在一个使用该线性硬件机构的实施方式中,该线性硬件机构使得拔出的第一刀片超过第一机箱中的其他刀片而从第一机箱突出离开,其中突出的刀片提供了刀片机箱中的哪个刀片已经物理地被拔出的可视指示符。

[0039] 现在参考图 5,为用于如上所述解耦合刀片的第四方法提供附加细节。考虑具有第一端 504 和第二端 506 的刀片 502。第二端 506 包括与机箱干线 510 中的物理插口 508 相配合的耦合销/槽/接头(例如,图 1 至图 2 所示的机箱干线 157、218 和/或 220)。刀片 502 安装在线性硬件机构 512 内,线性硬件机构 512 包括为这里所述的线性运动提供支撑而固定安装的线性致动器 514。线性致动器可以是液压致动器、齿轨和小齿轮,或能够产生足够线性力以接合/释放第二端 506 和物理插口 508 的任何其他机构。当刀片 502 物理地释放于具有其他刀片的机箱,线性运动使得刀片 502 从机箱干线 510 突出离开,从而给出了关于相比于其他刀片 516 哪个刀片已经从机箱释放(并且因此应该针对缺陷、工作分布问题、异常等进行检验)的快速可视提示,其他刀片 516 由于耦合至机箱干线 510 而在物理上接近于机箱干线 510。该物理释放防止刀片无意地被允许在机箱内保持活动。

[0040] 现在返回图 4,一旦已经将负担过重的第一刀片从第一机箱释放,则将第二机箱中的第二刀片接合到第二机箱中。该第二刀片可以是这样一种刀片,该第二刀片之前没有耦合至第二机箱,并且现在可以相反地使用用于释放第一机箱中的负担过重的第一刀片的上述四种方法中任何一种来进行接合。备选地,第二刀片可以仅是未充分使用的刀片,如由用于识别第一机箱中的负担过重的第一刀片的上述步骤确定的。在两个场景之一中,在本发明的优选实施方式中,第二刀片没有引导操作系统(如果其根本没有耦合至第二机箱)或

具有错误的操作系统（如果其之前耦合至第二机箱）。因此，在将原本应该发送到第一机箱中的负担过重的第一刀片的工作分配给第二机箱中的第二刀片之前，利用关闭第一刀片之前正在运行的同一操作系统的副本来引导第二刀片（块 424）。此后，原本应该发送到第一机箱中的第一刀片的任何未来工作现在由负载平衡管理器发送到第二机箱中的第二刀片（块 46）。尽管示出该处理结束于终止块 428，但是应该理解，如果第二机箱变得负担过重，则该处理可以重新迭代，因此使得失效备援回到同一第一刀片或第一机箱中的另一刀片。

[0041] 现在参考图 6，示出了本发明的备选实施方式，并且尤其示出了如何禁用第一刀片以及使能第二刀片（参见图 4 中的块 420 和 422）。初始地，第一机箱中的第一刀片（示出为刀片 602）经由机械耦合器 606 耦合至其专用硬盘驱动器 604（例如，图 1 所示的硬盘驱动器 134）。机械耦合器 606（包括处于刀片耦合软件 610 控制之下的处理逻辑 608）包括将刀片 602 机械耦合至硬盘驱动器 604 的物理开关。如果刀片耦合软件 610 检测到刀片 602 已经负担过重（根据上述参数），那么指示机械耦合器 606 将刀片 602 从硬盘驱动器 604 机械地断开，并且将刀片 612 机械地耦合至硬盘驱动器 604。刀片 612 可以经由诸如 LAN 等网络 614 耦合至机械耦合器 606。备选地，刀片 612 可以直接以及机械地连接至机械耦合器 606。在优选实施方式中，刀片 612 处于与刀片 602 不同的另一刀片机箱中（例如，刀片 602 是刀片 214n，而刀片 612 是图 2 所示的刀片 216a）。备选地，刀片 602 和刀片 612 均处于同一刀片机箱内。通过将第二刀片 612 耦合至硬盘驱动器 604，而将第一刀片 602 从硬盘驱动器 604 解耦合，现在第一刀片 602 不能读取去往和来自硬盘驱动器 604 的数据和指令，因此禁用它。此外，通过将第二刀片 612 耦合至硬盘驱动器 604，第二刀片 612 现在能够 1) 引导第一刀片的操作系统，并且 2) 读取之前由第一刀片使用的环境数据和应用软件，从而使第二刀片能够无缝地承担第一刀片的角色和责任。

[0042] 应该理解，本发明的至少一些方面可以备选地以包含程序产品的计算机可读介质来实现。定义本发明功能的程序可以经由各种有形的信号承载介质被递送到数据存储系统或计算机系统，这些信号承载介质包括但不限于：非可写存储介质（例如，CD-ROM）以及可写存储介质（例如，硬盘驱动器、读/写 CD ROM、光介质）。因此，应该理解，当此类存储介质由计算机可读指令编码时，其表示本发明的备选实施方式，其中该计算机可读指令指导本发明的方法功能。此外，应该理解，本发明可以由如下系统实现，该系统具有硬件、软件或如上所述的软件和硬件的组合或其等同物形式的装置。

[0043] 本文所使用的术语仅出于描述特定实施方式的目的，并且不旨在对本发明进行限制。如这里所用，除非上下文明确指出，单数形式“一个”、“一种”和“该”也旨在包括多个形式。还应该理解，当在说明书中使用术语“包括”和/或“包含”时，其指出了所述特征、整体、步骤、操作、元素和/或组件的存在，但不排除存在或添加一个或多个其他特征、整体、步骤、操作、元素、组件和/或其群组。类似地，术语“机箱”和“刀片机箱”应理解为可互换，可互换的术语还有“刀片”和“服务器刀片”、“机箱干线”和“干线”以及其上下文应该被理解为与类似术语同义的其他术语。

[0044] 以下权利要求书中的所有装置或步骤加功能元件的相应结构、材料、动作以及等同物旨在包括用于执行与如特别要求保护的其他要求保护的元件组合的功能的任何结构、材料或动作。已经出于示出以及描述的目的示出了本发明的描述，但是其不旨在穷举本发明或将本发明限于公开的形式。在不释放本发明的范围和精神的情况下，对于本领域的技

术人员而言很多修改和变形将变得明显。为了最好地解释本发明的原理和实际应用,并且为了使本领域的技术人员能够针对具有适于所构思的特定用途的各种修改的各种实施方式理解本发明,选择并描述了实施方式。

[0045] 应该进一步指出,如上述示例性实施方式中所描述的、本公开内容中所描述的方法可以通过使用 VHDL(VHSIC 硬件描述语言)程序和 VHDL 芯片来实现。VHDL 是一种示例性设计输入语言,其用于现场可编程门阵列(FPGA)、专用集成电路(ASIC)和其他类似的电子设备。因此,本文描述的任何软件实现的方法可以由基于硬件的 VHDL 程序来仿真,继而其可以应用于 VHDL 芯片,诸如 FPGA。将 VHDL 指令应用于 VHDL 芯片不仅引起 VHDL 芯片的物理转变,而且此类 VHDL 指令应用还可以使得特定加载的 VHDL 芯片重新耦合(物理上和/或逻辑上)至计算机系统内的其他硬件,因此引起附加的计算机系统的物理转变。

[0046] 因此,已经详细地并且通过参考其说明性实施方式描述了本申请的发明,所以在不释放所附权利要求书中限定的本发明的范围的情况下,修改和变形也是可能的。

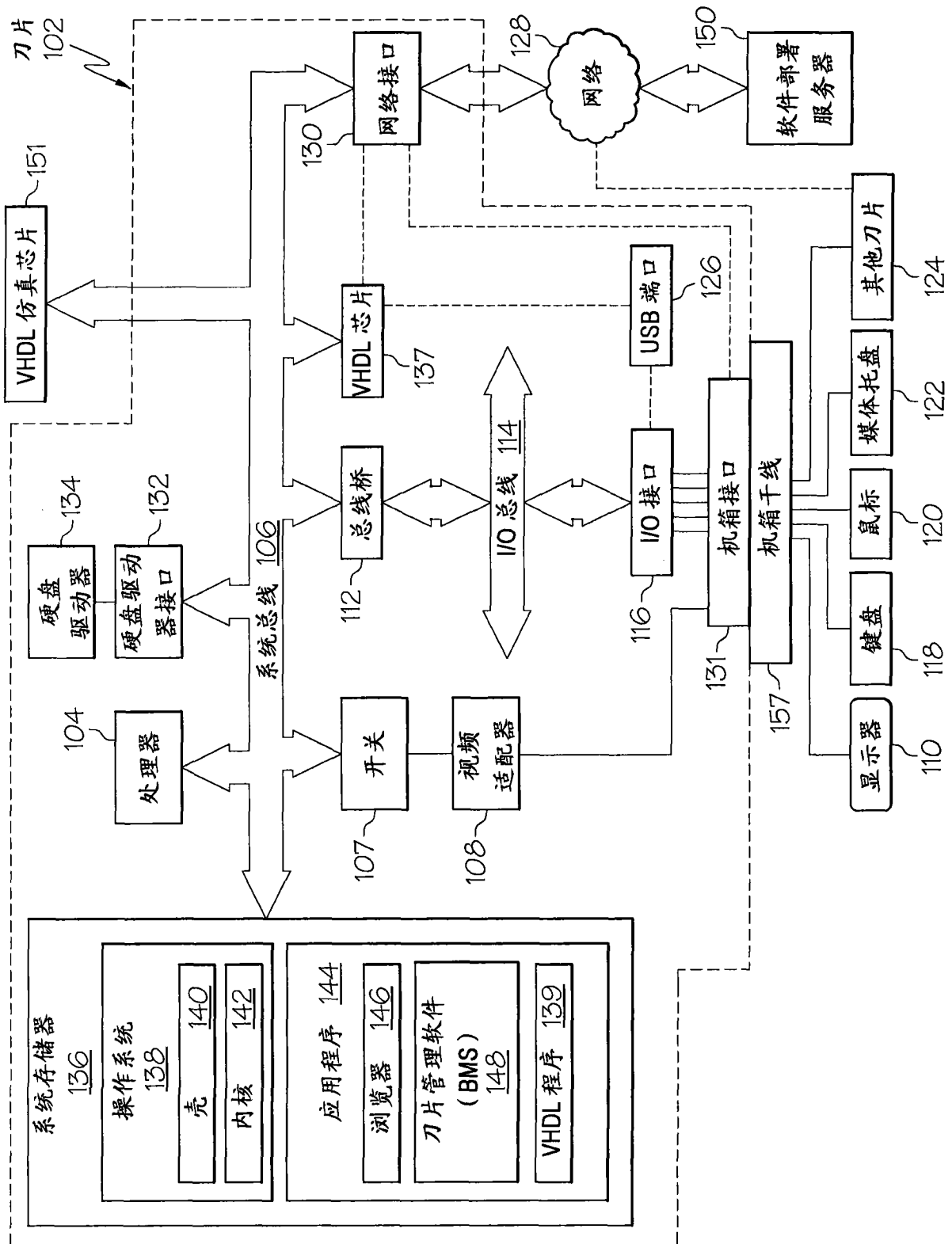


图 1

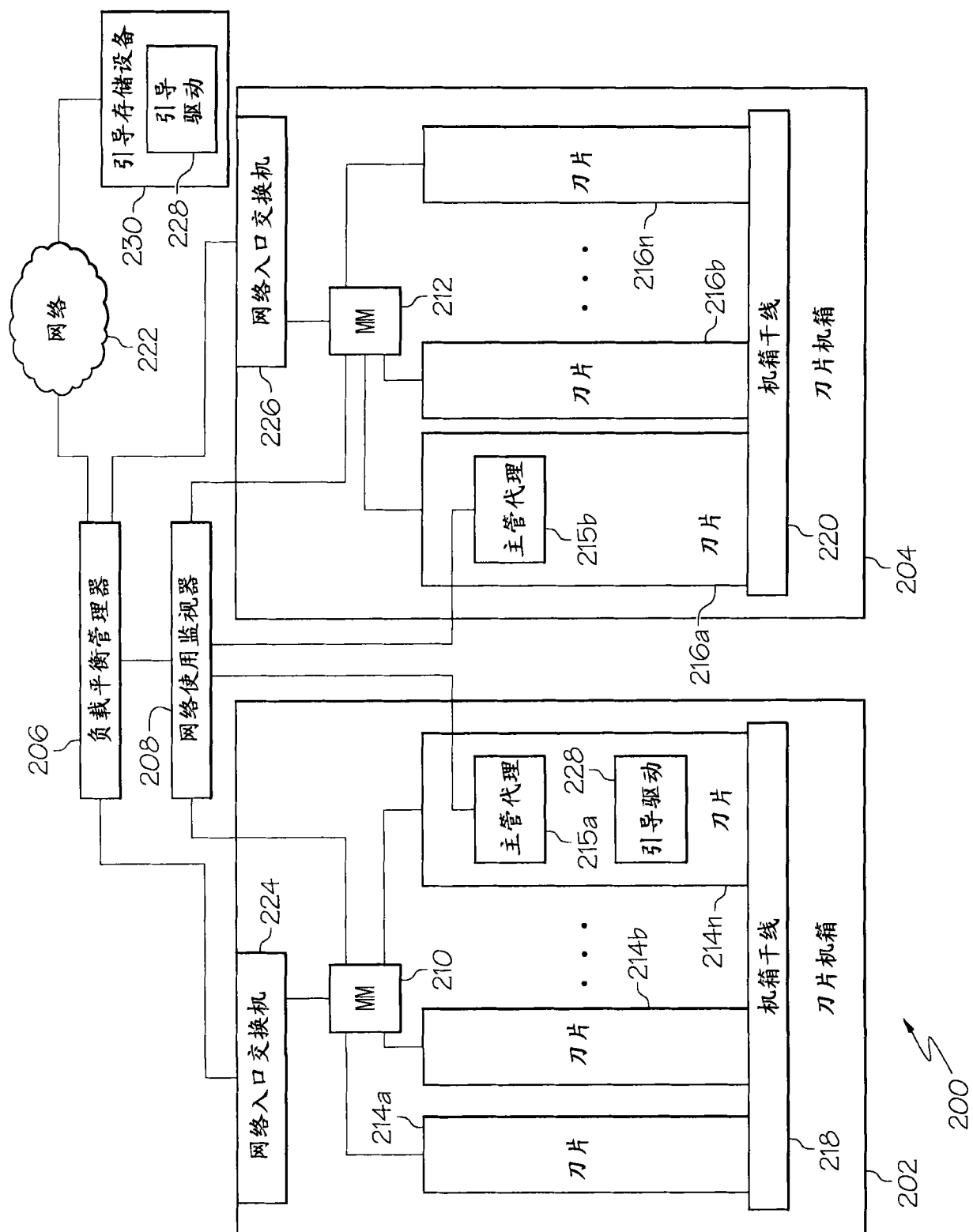


图 2

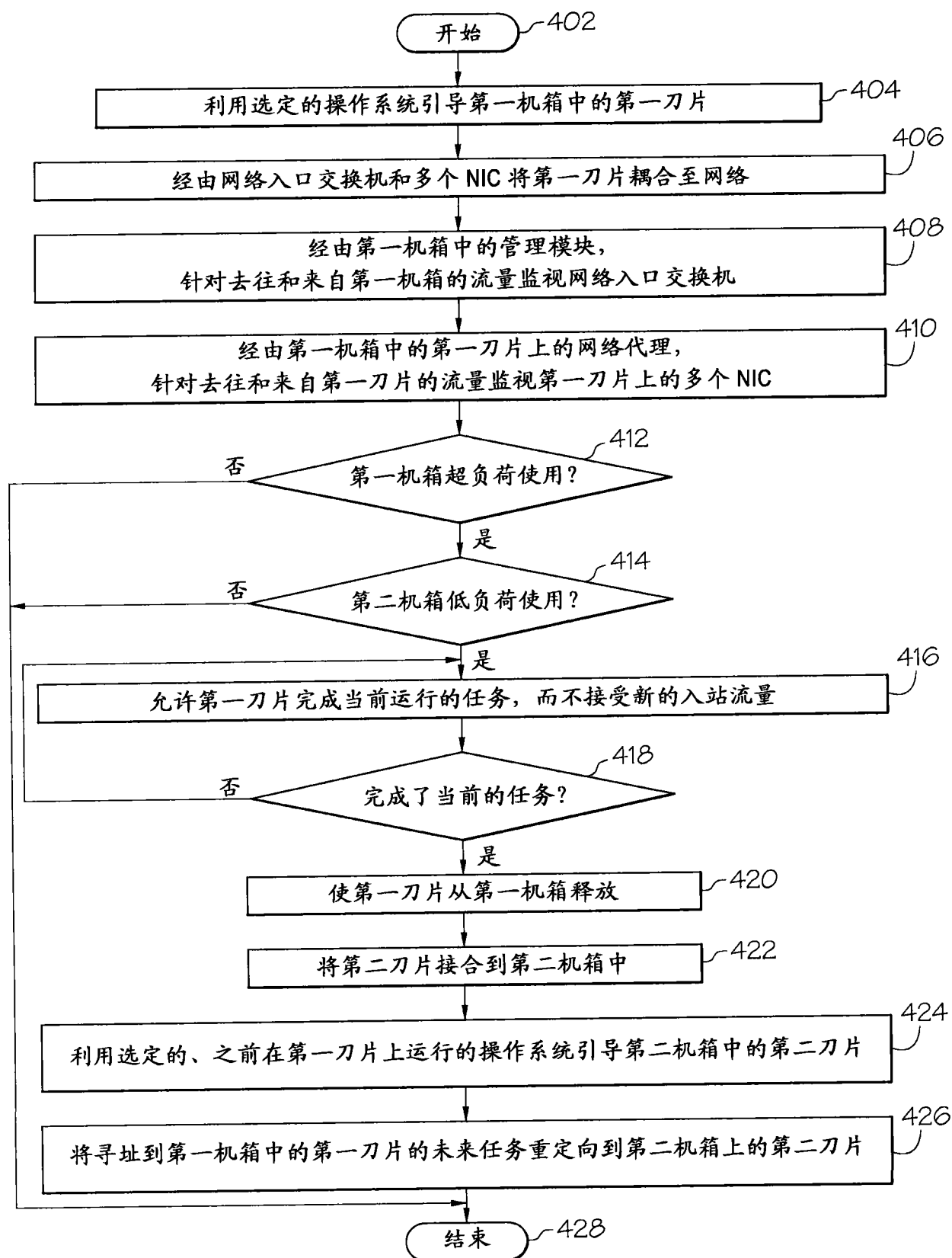


图 4

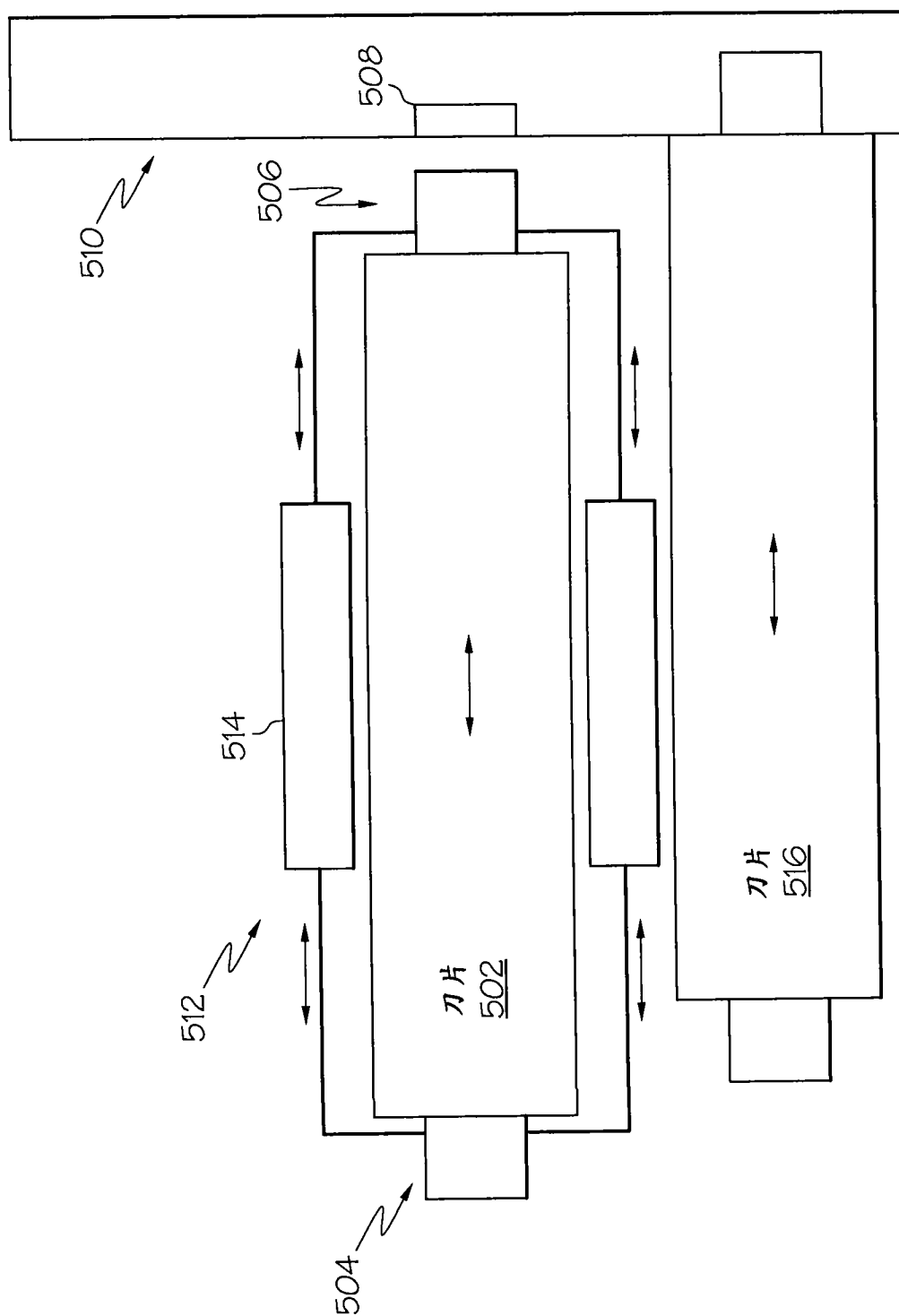


图 5

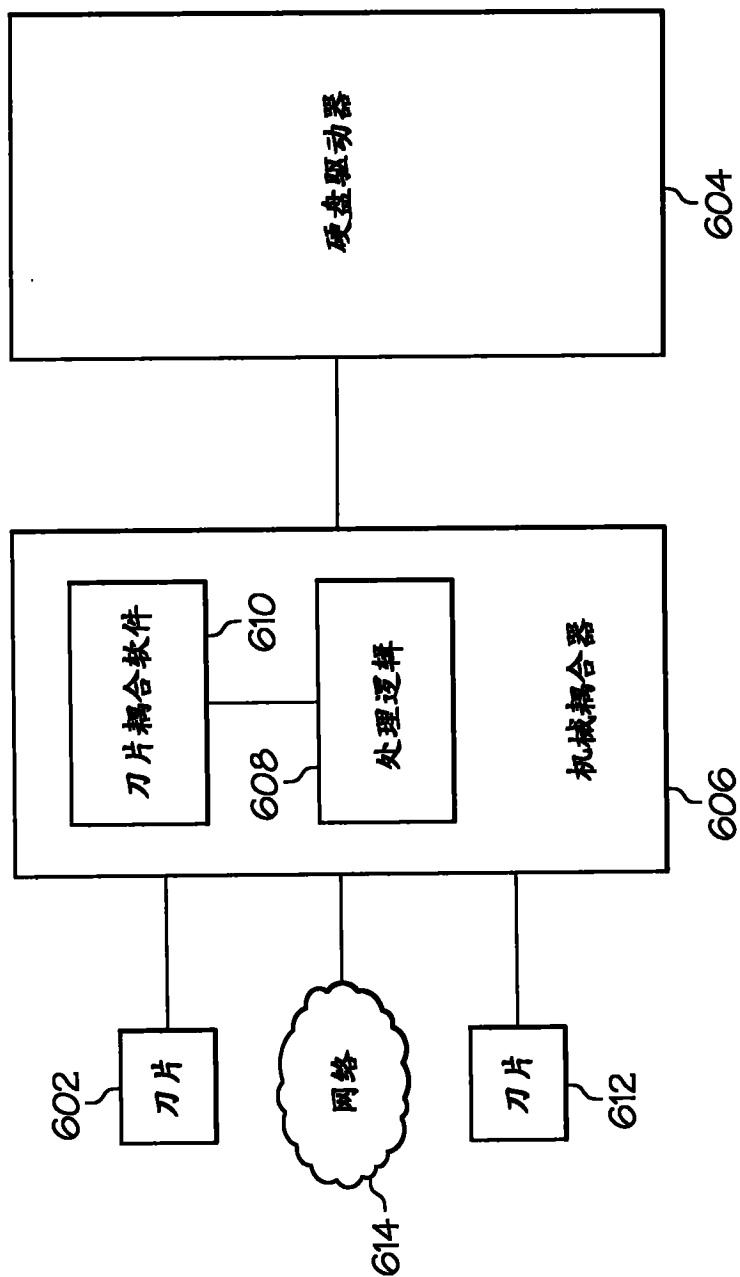


图 6