(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112088365 B
(45)授权公告日 2024.10.18

(21)申请号 201980029897.1
(22)申请日 2019.03.22
(65)同一申请的已公布的文献号
申请公布号 CN 112088365 A
(43)申请公布日 2020.12.15
(30)优先权数据
62/647,335 2018.03.23 US
62/690,533 2018.06.27 US
(85)PCT国际申请进入国家阶段日
2020.11.03
(86)PCT国际申请的申请数据
PCT/US2019/023712 2019.03.22
(87)PCT国际申请的公布数据
WO2019/183566 EN 2019.09.26

(73)专利权人 视界软件公司
地址 美国北卡罗莱纳州
(72)发明人 蒂莫西·A.·马丁
迈伦·L.·斯图尔特
(74)专利代理机构 北京高沃律师事务所 11569
专利代理师 王立普

(51)Int.Cl.
G06F 11/30(2006.01)
(56)对比文件
US 2011219118 A1,2011.09.08
US 2014075380 A1,2014.03.13
审查员 孟贵宇

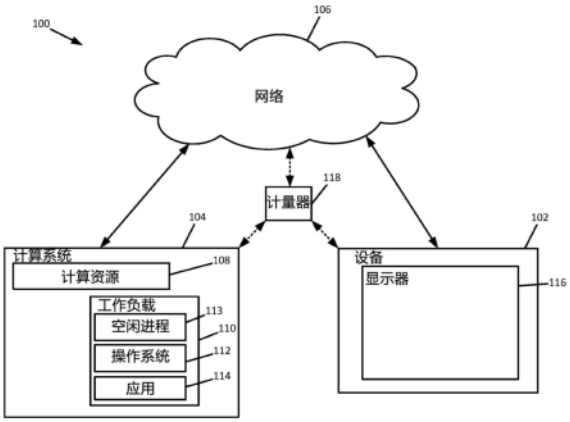
权利要求书2页 说明书26页 附图10页

(54)发明名称

将不同计算资源的使用量化为单个测量单元

(57)摘要

一种用于量化资源使用的方法可以包括测量在一段时间内正在被使用的多个物理计算资源的量。该方法还可以包括归一化正在被使用的每个物理计算资源的每个测量的量。该方法还可以包括对正在被使用的物理计算资源的归一化的测量的量进行求和,以生成表示在该时间段内正在被使用的物理计算资源的单个使用值。



CN 112088365 B

1.一种量化计算机资源的方法,包括:

接收多个物理计算资源中的每一个物理计算资源的一个或更多个资源属性,所述多个物理计算资源包括至少一个图形处理系统,所述图形处理系统具有作为资源属性的(i)图形处理器可变时钟速度和(ii)可变图形存储器时钟速度;

在一段时间内接收所述多个物理计算资源中的每一个物理计算资源的资源计算部分和资源分配;

通过以下步骤确定所述多个物理计算资源在所述一段时间内的单个使用值:

至少基于所述多个物理计算资源中的每个相应物理计算资源的一个或多个资源属性,使用一个或更多个处理器来计算所述多个物理计算资源中的每一个物理计算资源的利用率;

针对所述多个物理计算资源中的每个相应物理计算资源,并且至少基于每个相应物理计算资源的资源分配和资源计算部分,生成转换单元;

针对所述多个物理计算资源中的每个相应物理计算资源,并至少使用与所述相应物理计算资源相关联的转换单元来归一化与所述相应物理计算资源相关联的利用率;以及

对所述多个物理计算资源中的每一个物理计算资源的归一化的利用率进行求和,以生成表示在所述时间段内所述多个物理计算资源的所述单个使用值。

2.根据权利要求1所述的方法,其中,所述多个物理计算资源包括多个不同物理计算资源。

3.根据权利要求1所述的方法,其中,所述多个物理计算资源还包括:现场可编程门阵列、专用集成电路、片上系统、数字信号处理器、微控制器或自适应计算加速平台中的一个或更多个。

4.根据权利要求1所述的方法,其中,对每个相应物理计算资源的利用率进行归一化还包括将相应物理计算资源的利用率除以与所述相应物理计算资源相关联的转换单元,所述转换单元是由所述相应物理计算资源的资源计算部分划分的资源分配。

5.根据权利要求1所述的方法,还包括:

为所述多个物理计算资源中的每一个可用物理计算资源确定一个或更多个可用资源属性;

确定表示在所述时间段内所述多个物理计算资源中的最大可用资源的单个可用资源值;

使用所述一个或更多个处理器,至少基于所述一个或更多个可用资源属性来计算所述多个物理计算资源中的每个可用物理计算资源的可用性利用率;

针对每个相应可用物理计算资源并且至少使用与所述相应可用物理计算资源相关联的转换单元归一化与所述相应可用物理计算资源相关联的可用性利用率;以及

对所述可用物理计算资源中的每一个可用物理计算资源的归一化的可用性利用率进行求和,以生成所述单个可用资源值。

6.至少一种计算机可读储存介质,所述至少一种计算机可读储存介质具有储存在所述至少一种计算机可读储存介质上的一个或更多个指令,当一个或更多个处理器执行所述一个或更多个指令时,所述一个或更多个指令使得一个或更多个操作被执行,所述操作包括:

接收多个物理计算资源中的每一个物理计算资源的一个或更多个资源属性,所述多个

物理计算资源包括至少一个图形处理系统,所述图形处理系统具有作为资源属性的(i)图形处理器可变时钟速度和(ii)可变图形存储器时钟速度;

在一段时间内接收所述多个物理计算资源中的每一个物理计算资源的资源计算部分和资源分配;

通过以下步骤确定所述多个物理计算资源在所述一段时间内的单个使用值:

至少基于所述多个物理计算资源中的每个相应物理计算资源的一个或多个资源属性,使用所述一个或更多个处理器来计算所述多个物理计算资源中的每一个物理计算资源的利用率;

针对所述多个物理计算资源中的每个相应物理计算资源,并且至少基于每个相应物理计算资源的资源分配和资源计算部分,生成转换单元;

针对所述多个物理计算资源中的每个相应物理计算资源,并至少使用与所述相应物理计算资源相关联的转换单元来归一化与所述相应物理计算资源相关联的利用率;以及

对所述多个物理计算资源中的每一个物理计算资源的归一化的利用率进行求和,以生成表示在所述时间段内所述多个物理计算资源的所述单个使用值。

7.根据权利要求6所述的计算机可读储存介质,其中,所述多个物理计算资源包括多个不同物理计算资源。

8.根据权利要求6所述的计算机可读储存介质,其中,所述多个物理计算资源还包括:现场可编程门阵列、专用集成电路、片上系统、数字信号处理器、微控制器或自适应计算加速平台中的一个或更多个。

9.根据权利要求6所述的计算机可读储存介质,其中,对每个相应物理计算资源的利用率进行归一化还包括将相应物理计算资源的利用率除以与所述相应物理计算资源相关联的转换单元,所述转换单元是由所述相应物理计算资源的资源计算部分划分的资源分配。

10.根据权利要求6所述的计算机可读储存介质,其中,所述操作还包括:

为所述多个物理计算资源中的每一个可用物理计算资源确定一个或更多个可用资源属性;

确定表示在所述时间段内所述多个物理计算资源中的最大可用资源的单个可用资源值;

使用所述一个或更多个处理器,至少基于所述一个或更多个可用资源属性来计算所述多个物理计算资源中的每个可用物理计算资源的可用性利用率;

针对每个相应可用物理计算资源并且至少使用与所述相应可用物理计算资源相关联的转换单元归一化与所述相应可用物理计算资源相关联的可用性利用率;以及

对所述可用物理计算资源中的每一个可用物理计算资源的归一化的可用性利用率进行求和,以生成所述单个可用资源值。

11.根据权利要求6所述的计算机可读储存介质,其中,所述经由一个或更多个插件接收所述多个物理计算资源中的至少一个物理计算资源的资源计算部分和资源分配。

12.根据权利要求6所述的计算机可读储存介质,其中,所述操作还包括确定与所述时间段内每个物理计算资源对应的高峰使用、低谷使用或平均使用中的一个或更多个。

将不同计算资源的使用量化为单个测量单元

[0001] 相关申请的交叉引用

[0002] 本申请要求在2018年3月23日提交的题为“将不同计算资源的使用量化为单个测量单元”的美国临时申请No.62/647,335和在2018年6月27日提交的题为“将不同计算资源的使用量化为单个测量单元”的美国临时申请No.62/690,533的权益,二者均通过引用完全并入本文。

技术领域

[0003] 本公开总体上涉及计算,并且更具体地涉及计量不同计算机资源的使用以及将使用量化为单个测量单元。

背景技术

[0004] 云计算允许用户(例如,个人或组织)租用由第三方维护和/或管理的计算资源。这可允许用户租用针对其个人计算需求定制的计算资源,而不必在物理硬件、数据中心房地产、电力、数据中心人员和/或软件许可证方面做出大量投资。

[0005] 示例云计算系统可以包括物理资源(例如,处理器和存储器)和/或虚拟化的资源(例如,虚拟化处理器和虚拟化存储器)。访问云计算系统的用户可租用能够运行其工作负载的大量资源。所租用的计算资源的量可基于理论最大工作负载来估计。假设理论最大工作负载使用几乎所有(例如,大约100%)可用资源。换句话说,用户将租用执行其工作负载所需的最大量的资源,以便避免由于资源需求的激增而导致的系统性能问题。然而,使用中的激增可能偶尔发生,并且平均使用的资源可能低于理论最大值,这导致浪费(例如,租用未使用的计算资源)。

[0006] 云计算提供商可基于与用户使用(或消耗)的实际资源量相对的租用(例如,预定时间)的资源量来实施定价模型。这些资源可被封装为具有预配置的中央处理单元(CPU)、存储器(例如,随机存取存储器)、储存器、图形处理单元(GPU)、(例如,提供预定联网速度的)专用联网能力、现场可编程门阵列(FPGA)和/或任何其他计算资源的服务器。所封装的服务器一般可被称为用户可租用一段时间(例如,以秒、分钟、小时、天、周、月、年和/或任何其他时间间隔的间隔)的实体。因此,用户可基于他们估计使用的最大计算资源(例如,其可基于使用中的激增来估计)来租用实体。通过租用实体中的资源,用户可在租赁协议的大部分期限内仅使用计算资源的一部分。换句话说,这可能导致用户支付未被消耗的资源并且云计算提供商可能由于租用未使用的资源而在财务上受益。

发明内容

附图说明

[0007] 通过阅读以下结合附图的详细描述,将更好地理解这些以及其他特征和优点,其中:

[0008] 图1示出了与本公开的实施例一致的计算网络的示意性示例。

[0009] 图2A示出了与本公开的实施例一致的能够在图1的计算网络中使用的计算系统的示意性示例。

[0010] 图2B示出了与本公开的实施例一致的图2A的计算系统的示意性示例。

[0011] 图3A示出了与本公开的实施例一致的能够在图2的计算系统中使用的用于测量计算资源的使用的计量器的示意性示例。

[0012] 图3B示出了与本公开的实施例一致的能够与图3A的计量器一起使用的度量/属性收集器的示意性示例。

[0013] 图3C示出了与本公开的实施例一致的能够与图3A的计量器一起使用的分析平台的示意性示例。

[0014] 图3D示出了与本公开的实施例一致的图示资源使用的绘图(或曲线图)。

[0015] 图4示出了与本公开的实施例一致的可以与图3的计量器一起使用的转换单元生成器的示意性示例。

[0016] 图5示出了与本公开的实施例一致的将不同计算资源的使用量化为单个测量单元的方法的示例流程图。

[0017] 图6示出了与本公开的实施例一致的将总的不同可用计算资源(例如,所使用资源和非活动资源两者)量化为单个测量单元的方法的示例流程图。

具体实施方式

[0018] 本文总体上公开了一种用于在计算和量化使用为单个测量单元时测量物理资源使用的系统、方法和装置。例如,可测量、归一化(例如,转换)不同计算资源到通用测量单元,并且可对归一化测量值中的每一个求和,使得总使用可表示为单个值。通过将不同计算资源表示为单个单元,可以实时监控资源使用(或消耗),可以优化工作负载以改进可用资源的使用,和/或可以减少浪费。

[0019] 通过测量物理资源(与虚拟资源相对),可获得实际使用的更准确的测量。例如,由于所使用的技术被称为膨胀,基于虚拟资源来测量使用可能高估物理资源的实际使用。换句话说,当仅测量虚拟资源时,虚拟资源可指示工作负载正在消耗(或使用)的资源多于工作负载在物理资源中实际消耗的资源。在一些情况下,被指示为正在被消耗的虚拟资源的量可能超过可用的物理资源。

[0020] 如本文中所使用的图形处理单元(GPU)可以泛指至少包括图形处理器和/或图形存储器的计算资源,并且在一些情况下,GPU可以泛指图形处理系统(GPS)。

[0021] 图1示出了计算网络100的示意性示例。如图所示,计算网络100可包括设备102和计算系统104,设备102和计算系统104均通信地耦接到网络106(例如,互联网)。计算系统104可包括被配置成执行一个或更多个工作负载110的物理计算资源108。一个或更多个工作负载110可以包括例如一个或更多个操作系统112、一个或更多个应用114和/或一个或更多个空闲进程113中的一个或更多个。空闲进程113可以是与计算系统104的操作相关的后台进程(例如,作为计算系统104的硬件的操作的结果)并且可以消耗物理计算资源108的至少一部分,即使还未接收到来自例如将要执行应用114的设备102和/或操作系统112的请求。换句话说,即使当计算系统104空闲时,由于当计算系统104上电时计算系统104的物理

硬件的操作,计算系统104仍然消耗物理计算资源108的至少一部分。

[0022] 响应于由设备102(例如,(多个)操作系统112和/或(多个)应用114)生成的请求而引起执行工作负载110中的一个或更多个。例如,设备102可以通过网络106将请求传送到计算系统104来请求一个或更多个应用114的执行。当接收到请求时,计算系统104可以将物理计算资源108的至少一部分分配给应用114的执行。响应于应用114被执行,计算系统104可将数据传输回设备102。例如,传输回设备102的数据可以被用于生成设备102的显示器116上的图形用户接口。然后,设备102的用户可以与图形用户接口交互,并且使得设备102向计算系统104传输附加指令,该附加指令使得在计算系统104上执行应用114的不同部分。

[0023] 如图所示,可以设置计量器118。计量器118可包括硬件、软件和/或固件的任何组合,该硬件、软件和/或固件的任何组合被配置成测量在一段时间(例如,预定或非预定的时间段)内由工作负载110(例如,空闲进程113、(多个)操作系统112和/或(多个)应用114)使用的物理计算资源108的使用。例如,计量器118可以被实施为软件,该软件被配置为储存在一个或更多个存储器(例如,任何类型的有形的、非暂时性的储存介质,以及可以包括磁记录介质(例如,硬盘驱动器)、光盘、半导体器件(如只读存储器(ROM))、随机存取存储器(RAM)(如动态和静态RAM)、可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、闪存、磁卡或光卡、或用于储存电子指令的任何类型的储存媒介中的一个或更多个)上并且由一个或更多个处理器(例如,由英特尔(INTEL)、高级微设备公司(ADVANCED MICRO DEVICES),IBM、ARM、甲骨文(ORACLE)提供销售的处理器和/或任何其他处理器)执行以便引起一个或更多个操作被执行。通过进一步的示例,计量器118可以被实施为电路(例如,专用集成电路)。

[0024] 物理计算资源108的使用可使用在计算系统104上执行的主机操作系统(例如,以汇编语言编写的操作系统)和/或管理程序来测量。例如,计量器118可以包括具有一个或更多个插件的插件数据库,所述插件被配置成使得能够经由在计算系统104上执行的管理程序来测量主机操作系统的资源使用。作为另一示例,计量器118可被配置以(例如,使用主机操作系统)测量直接来自物理计算资源108的资源使用。在一些情况下,计量器118可跨多个不同的计算系统104来测量物理计算资源使用。

[0025] 计量器118测量物理计算资源108的使用,将每个物理计算资源108归一化为通用测量单元,并且组合经归一化的计算资源,使得所组合的(例如,总的)使用可被表示为单个单元。换句话说,计量器118可通常被描述为将多个不同计算资源的使用量化为表示计算资源108的使用的单个值。例如,计量器118可被配置成在一段时间内测量多个不同计算资源的使用并且生成表示在该时间段内的使用的单个使用值。

[0026] 为了减轻和/或以其他方式防止膨胀的影响,计量器118测量物理资源使用而不是例如虚拟资源使用。这样,可更准确地测量例如虚拟机器的物理资源使用。在一些情况下,计量器118还可以被配置成测量被使用的可用物理计算资源108(例如,被使用的和不活动的资源)的最大量,并被配置成将所测量的物理计算资源108的每个物理资源的量归一化为通用测量单元,然后对归一化的资源求和以获得表示为单个单元的可用的物理资源的最大量。通过测量可用的资源的最大量,用户能够优化用于执行工作负载110的物理计算资源108的量。另外或替代地,计量器118可被配置成测量运行计算系统104所需的物理计算资源108(例如,仅由空闲进程113所使用的资源)的最小量,将所测量的物理计算资源108的每个

物理资源的量归一化为通用测量单元,然后对归一化的资源求和以获得表示需要(例如,当计算系统104正在操作但不执行一个或更多个操作系统112和/或应用114时)使用的物理资源的最小量的单个单元。

[0027] 在一些情况下,计量器118可被配置成测量一段时间内(例如,资源使用的计费期、合同条款、小时、天、周和/或任何其他时间段)的物理资源的使用并且生成表示该时间段内的物理资源的使用的单个使用值。然后,所测量的使用可以用于确定在一段时间内要购买的资源的适当量和/或用于基于实际使用生成账单。因此,资源消耗用户可以减少购买的资源量,并且资源提供商可以减少提供用于销售的资源的总量和/或将未使用的资源分配给另一个用户。

[0028] 在一些情况下,计量器118可以被配置成启用和/或禁用设备102对计算系统104的计算资源108的访问。换句话说,计量器118可以被配置成选择性地将设备102通信地耦接和/或解耦到计算系统104。例如,计量器118可以在一段时间期满(例如,合同条款的终止)之后、未能满足合同条款(例如,未能支付账单、滥用计算资源等)、和/或响应于一段时间内测量的资源使用超过或未能超过预定的量(或阈值)禁用设备102对计算资源108的至少一部分的访问。作为另一示例,计量器118可以使得能够在预定时间(例如,当合同条款开始时)访问计算资源108。

[0029] 在一些情况下,计量器118可被配置成基于可用性选择性地启用和禁用对计算资源108的至少一部分的访问。例如,当另一用户未使用计算资源108时,设备102可访问计算资源108的至少一部分。然而,当另一个用户请求访问计算资源108时,召回请求可以被传输到计量器118,该召回请求使得计量器118禁用设备102对计算资源108的至少一部分的访问,使得另一个用户可以访问计算资源108。因此,计量器118通常可以被描述为被配置成减少未被用于例如执行一个或更多个应用114或操作系统112的计算资源108的量。

[0030] 如图所示,计量器118可以通信地耦接到设备102、计算系统104和/或网络106。在一些情况下,计量器118的至少一部分可以被包括在设备102和/或计算系统104中。附加地或替代地,计量器118的至少一部分可以被包括在(例如,使用网络106)通信地耦接到设备102中的一个或更多个和/或计算系统104的第三方设备中。

[0031] 设备102可以包括个人计算机、平板计算机、手机、智能电话、智能手表、智能电视、健身追踪器、智能秤(用于称重物体)、智能恒温器、智能安全监控系统(例如,室内和/或室外摄像机系统、门铃、警报系统、锁系统、照明系统和/或任何其他智能安全监控系统)、智能显示器(例如,被配置为与网络106通信的显示器)、智能医疗保健系统(例如,超声系统、分子成像系统、计算机化的断层扫描系统、X射线系统和/或任何其他智能医疗保健系统),远程健康监控设备、游戏设备(例如游戏机、手持游戏设备、虚拟现实耳机和/或任何其他游戏设备)、物联网设备(例如自动驾驶车辆、连接的家用电器、无人机、机器人和/或任何其他物联网设备)、服务器和/或任何其他能够与计算系统104通信的设备中的任何一个或更多个。

[0032] 图2A示出了图1的计算系统104的示意性示例。如图所示,计算系统104的物理计算资源108可包括中央处理单元(CPU)202、存储器204、储存器206、网络接口208和图形处理单元(GPU)210。CPU202、存储器204、储存器206、网络接口208和GPU210中的每一个都具有与其相关联的一个或更多个度量/属性。

[0033] 附加地或可替代地,计算资源108可包括其他资源209。其他资源209可包括例如现

场可编程门阵列(FPGA)、张量处理单元(TPU)、智能处理单元(IPU)、神经处理单元(NPU),视频处理单元(VPU)、数字信号处理器(DSP)、加速处理单元(APU),专用集成电路(ASIC)、片上系统(SoC)、可编程SoC、专用标准部件(ASSP),自适应计算加速平台(ACAP)、微控制器和/或任何其他计算资源中的任何一个或更多个。

[0034] 如图2B中所示,CPU 202可以与用于多个处理核212、处理器时钟速度214和/或处理器负载216的量度/属性相关联。处理核212的数量指示CPU 202可用的独立处理单元(或核)的数量。处理器时钟速度214指示用于CPU 202的核中的每个核的默认时钟速度。处理器负载216指示由CPU 202执行的计算工作的量。CPU 202可以是任何计算机处理器,包括例如能够执行计算机指令的单核和/或多核处理器。CPU 202的示例可以包括由英特尔、高级微设备公司、IBM、ARM、甲骨文提供销售的处理器和/或任何其他处理器。

[0035] 存储器204可与可用存储器的总量218和/或所使用的存储器的量220的度量/属性相关联。存储器204可以是有形的、非暂时性的存储器。存储器204可以是易失性存储器。例如,存储器204可以是随机存取存储器(RAM)。RAM的示例可以包括例如静态RAM(SRAM)、动态RAM(DRAM)、快速页面模式的DRAM(FPM DRAM)、扩展数据输出DRAM(EDO DRAM)、同步动态RAM(SDRAM)、双倍数据速率SDRAM(DDR SDRAM)、RAMBUS动态RAM(RDRAM)和/或任何其他形式的RAM。

[0036] 储存器206可以与用于以下各项的度量/属性相关联:可用的储存器的总量222、所使用的储存器的总量224、写入到储存器206的字节数226和/或从储存器206读取的字节数228。储存器206可以是非易失性存储器。例如,储存器206可以包括任何类型的有形的、非暂时性存储介质,以及可包括磁性记录介质(例如,硬盘驱动器)、光盘、诸如只读存储器(ROM)的半导体器件、诸如动态和静态RAM的随机存取存储器(RAM),可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、闪存,磁卡或光卡,或用于储存电子指令的任何类型的储存媒介中的任何一个或更多个。

[0037] 网络接口208可与用于被发送的字节数230和/或被接收的字节数232的度量/属性相关联。网络接口208可配置为通信地耦接到网络106(图1)。例如,网络接口208可以是网络接口控制器,该网络接口控制器配置为通信地耦接到以太网、无线网络、光纤通道网络、光纤分布式数据接口(FDDI)网络、铜线分布式数据接口(CDDI)网络和/或任何其他网络中的一个或更多个。

[0038] GPU 210可与用于以下各项的度量/属性相关联:图形处理器默认时钟速度234,图形处理器可变时钟速度236、图形处理器负载238、图形存储器默认时钟速度240,可变图形存储器时钟速度242、图形存储器负载244、图形存储器总线246,可用的图形存储器的总量248、所使用的图形存储器的量250,和/或着色器252的量。图形处理器默认时钟速度234指示GPU 210的处理器的默认时钟速度,且图形处理器可变时钟速度236指示图形处理器的可变时钟速度。在一些情况下,图形处理器可变时钟速度236可超过图形处理器默认时钟速度234的时钟速度。在其他情况下,图形处理器可变时钟速度236可能不超过图形处理器默认时钟速度234。类似地,图形存储器默认时钟速度240指示图形存储器的默认时钟速度,并且可变图形存储器时钟速度242指示图形存储器实际运行的时钟速度。在一些情况下,可变图形存储器时钟速度242可测量大于图形存储器默认时钟速度240。在其他情况下,可变图形存储器时钟速度242可以测量小于(或等于)图形存储器默认时钟速度240。图形存储器负载

244指示在GPU存储器上执行的读取和/或写入操作的量。图形存储器总线246指示GPU210的总线大小。

[0039] GPU 210可以包括集成的和/或专用的GPU。例如,GPU可以包括由例如英特尔、NVIDIA、高级微设备公司制造的任何GPU和/或任何其他GPU。当GPU 210包括集成GPU时,度量/属性可以至少部分地基于集成GPU的共享资源。

[0040] 虽然CPU 202、存储器204、储存器206、网络接口208和GPU 210少有提及,但应理解,计算系统104可包含多个物理CPU、存储器、储存器、网络接口和GPU,其中的一个或更多个可配置成协作,以例如改进计算系统104的性能。类似地,在计算系统104包括其他资源209中的一个或更多个的情况下,可以存在其他资源209中的每个资源的一个或更多个。

[0041] 图3A示出了图1的计量器118的示意性示例。如图所示,计量器118可以包括度量/属性收集器302、计算器305和分析平台312,所述度量/属性收集器302通信地耦接到物理计算资源108(图2)。度量/属性收集器302、计算器305和/或分析平台312中的每一个都可以分布在不同的计算系统(或环境)上。在一些情况下,度量/属性收集器302、计算器305和/或分析平台312中的一个或更多个可以位于相同的计算系统上。例如,度量/属性收集器302、计算器305和/或分析平台312中的一个或更多个可以位于设备102上。作为进一步的示例,度量/属性收集器302、计算器305和/或分析平台312中的一个或更多个可以位于通信地耦接到网络106和/或计算系统104的单独设备上。作为又一示例,度量/属性收集器302、计算器305和/或分析平台312中的一个或更多个可以位于计算系统104上。

[0042] 度量/属性收集器302从CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个中的一个或更多个请求一个或更多个度量/属性。例如,度量/属性收集器302可以从在CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个上(例如,使用插件)执行的管理程序请求一个或更多个度量/属性。作为进一步的示例,度量/属性收集器302可以从在CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个中的一个或更多个上执行的主机操作系统(例如,以汇编语言编写的操作系统)请求一个或更多个度量/属性。

[0043] 当度量/属性收集器302从CPU 202请求一个或更多个度量/属性时,所请求的度量/属性可以被输入到CPU周期利用率计算器304中。CPU周期利用率计算器304可以至少部分地基于来自CPU 202的所请求的度量/属性来生成与正被利用的CPU周期的量相对应的度量/属性。例如,CPU周期利用率计算器304可以至少部分地基于处理核212的数量、处理器时钟速度214和处理器负载216来生成CPU周期利用率。在这些情况下,CPU周期利用率计算器304可以根据以下等式生成正在利用的CPU周期:

[0044] $$\frac{(\text{处理器时钟速度})(\text{处理核的数量})}{100} * (\text{处理器负载}) \qquad [\text{等式 1}]$$

[0045] 当度量/属性收集器302从GPU 210请求一个或更多个度量/属性时,所请求的度量/属性可以被输入到GPU处理器周期利用率计算器306和/或GPU存储器周期利用率计算器308中的一个或更多个。GPU处理器周期利用率计算器306和GPU存储器周期利用率计算器308可以至少部分地基于来自GPU 210的所请求的度量,生成分别与GPU处理器周期利用率和GPU存储器周期利用率相对应的度量/属性。

[0046] 例如,GPU处理器周期利用率计算器306可以至少部分地基于图形处理器可变时钟

速度236和图形处理器负载238来生成GPU处理器周期利用率。在这些情况下,GPU处理器周期利用率计算器306可以根据以下等式生成所利用的GPU处理器周期:

[0047] $$\frac{(\text{图形处理器可变时钟速度})}{100} * (\text{图形处理器负载}) \qquad [\text{等式 2}]$$

[0048] 作为进一步的示例,GPU存储器周期利用率计算器308可以至少部分地基于可变图形存储器时钟速度242和图形存储器负载244来生成GPU周期利用率。在这些情况下,GPU存储器周期利用率计算器308可以根据以下等式生成所消耗的GPU存储器的时钟周期:

[0049] $$\frac{(\text{可变图形存储器时钟速度})}{100} * (\text{图形存储器负载}) \qquad [\text{等式 3}]$$

[0050] 当度量/属性收集器302从储存器206请求一个或更多个度量/属性时,度量/属性可以被输入到磁盘IO计算器310中。磁盘IO计算器310可以生成与在一段时间(例如,一秒)内读取和写入到储存器206的比特的总数相对应的度量/属性。因此,在某些情况下,磁盘IO计算器310可以将该段时间内写入的字节数226和读取的字节数228相加,以生成对应于磁盘IO的度量/属性。在时间段大于一秒的情况下,写入的字节数226和读取的字节数228可以除以秒的总数。在时间段小于一秒的情况下,写入的字节数226和读取的字节数228可以除以一秒的分数。

[0051] 当度量/属性收集器302从网络接口208请求一个或更多个度量/属性时,度量/属性可以被输入到网络IO计算器311中。网络IO计算器311可以生成与网络接口208在一段时间(例如,一秒)内发送或接收的比特总数相对应的度量/属性。因此,在某些情况下,网络IO计算器311可将该段时间内发送的字节数230和接收的字节数232相加,以生成对应于网络IO的度量/属性。在时间段大于一秒的情况下,发送的字节数230和接收的字节数232可以除以秒的总数。在时间段小于一秒的情况下,发送的字节数230和接收的字节数232可以除以一秒的分数。

[0052] 如图所示,分析平台312可以接收与物理计算资源108相关联的度量/属性中的一个或更多个,包括例如从计算器305中的一个或更多个输出的一个或更多个度量/属性。在一些情况下,分析平台312可以包括与相应的其他资源209中的一个或更多个相对应的一个或更多个其他资源计算器331。从计算器305中的一个或更多个输出的一个或更多个度量/属性可以包括例如(例如,作为来自CPU周期利用率计算器304的输出的)CPU周期利用率、(例如,作为由GPU处理器周期利用率计算器306的输出的)GPU处理器周期利用率、(例如,作为由GPU存储器周期利用率计算器308的输出的)GPU存储器周期利用率、(例如,作为由磁盘IO计算器310的输出的)磁盘IO和/或(例如,作为由网络IO计算器311的输出的)网络IO中的一个或更多个。分析平台312可以使用一个或更多个转换单元来转换所接收的度量/属性。在一些情况下,每个度量/属性具有对应的转换单元。例如,分析平台312可以利用CPU转换单元314、存储器转换单元316、储存器转换单元318、网络转换单元320、GPU处理器转换单元322、GPU存储器转换单元324、GPU存储器速度转换单元326、磁盘IO转换单元328和/或对应于计算资源的任何其他转换单元(例如,用于诸如FPGA资源、ASIC资源、SoC资源、DSP资源、微控制器资源、ACAP资源等的其他资源209中的一个或更多个的转换单元)。

[0053] 分析平台312可以通过将每个物理资源(例如,CPU 202、存储器204、储存器206、网

络接口208和GPU 210)的使用除以与该物理资源相对应的相应转换单元(例如,CPU转换单元314、存储器转换单元316、储存器转换单元318、网络转换单元320、GPU处理器转换单元322、GPU存储器转换单元324、GPU存储器速度转换单元326和磁盘IO转换单元328)来将不同计算资源转换成归一化单元,以获得每个物理资源的相应归一化使用。在一些情况下,分析平台312可以被配置为将其他资源209中的一个或更多个转换为标准化单元。例如,分析平台312可以包括其他资源转换单元329。

[0054] 然后,可以对归一化使用进行求和以获得总使用值。另外或替代地,如鉴于本文的公开内容应当容易理解的,在一些情况下,可以根据计算资源类型(例如,CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个)对归一化使用进行求,以获得特定资源的归一化使用值。例如,可以对GPU 210的物理资源(例如,GPU处理器和/或GPU存储器)的归一化使用进行求和以生成总GPU使用。特定物理资源的归一化使用的其它示例可以包括总FPGA使用、总ASIC使用、总DSP使用、总SoC使用、总微控制器使用、总ACAP使用、总TPU使用和/或与相应资源相对应的任何其它总使用。

[0055] 图3B示出了度量/属性收集器303,度量/属性收集器303可以是度量/属性收集器302的示例,度量/属性收集器303被配置为监控许多个(例如,多个)不同用户的资源使用,并将每个用户的使用与相应的用户相关联。换句话说,度量/属性收集器303被配置为接收与资源使用相对应的度量。度量/属性收集器303可以包括硬件、软件和/或固件的任何组合。

[0056] 度量/属性收集器303可以被配置为与计算系统104、设备102和/或通信地耦接到网络106的任何其他设备中的任何一个或更多个包括在一起。例如,设备102可以包括度量/属性收集器303。当设备102包括度量/属性收集器303时,与不在设备102上执行度量/属性收集器303时相比,度量/属性收集器303的用户能够通过对用户的网络安全性的更少修改来使用计量器118。

[0057] 如图所示,度量/属性收集器303经由通信链路330通信地耦接到计算系统104。通信链路330可将数据(例如,与计算资源108相关联的度量/属性)传输到度量/属性收集器303。可以对通过通信链路330传输的数据进行加密,以防止对数据的未授权访问。如图所示,度量/属性收集器303可远离计算系统104(例如,在不同的服务器上、在不同的网络中、和/或在不同的设备上操作)。然而,在一些情况下,度量/属性收集器303可以在计算系统104本地(例如,在同一服务器上、在同一网络上和/或在同一设备上操作)。

[0058] 度量/属性收集器303还可以被配置为通过通信链路362与分析平台312通信。可以对在通信链路362上传输的数据进行加密以防止对数据的未授权访问。在一些情况下,度量/属性收集器303可以包括分析平台312。照此,度量/属性收集器303可以不在通信链路362上传输数据。

[0059] 分析平台312可以生成表示为单个归一化单元的总资源使用值。如图所示,分析平台312可远离度量/属性收集器303(例如,在不同的服务器上、在不同的网络中、和/或在不同的设备上操作)。然而,在一些情况下,分析平台312可以在度量/属性收集器303本地(例如,在同一服务器上、在同一网络上和/或在同一设备上操作)。

[0060] 如图所示,度量/属性收集器303可包括至少一个收集器核332,至少一个收集器度量接收器334、至少一个收集器同步器336、至少一个插件数据库338、至少一个收集器自动

器340、至少一个机器管理器342、至少一个收集器数据库管理器344,至少一个账户管理器346和至少一个机器集合管理器348。收集器核332与收集器度量接收器334、收集器同步器336、插件数据库338、收集器自动器340、机器管理器342、收集器数据库管理器344、账户管理器346和机器集合管理器348中的每一个进行通信和/或管理它们。

[0061] 收集器度量接收器334被配置为接收表示一个或更多个用户和/或工作负载的计算资源108的使用的度量/属性。在一些情况下,收集器度量接收器334可包括计算器305中的一个或更多个。

[0062] 资源使用的至少一部分可与一个或更多个用户相关联。例如,使用账户管理器346可将资源使用的至少一部分与一个或更多个账户350相关联。账户350可以与一个或更多个资源消耗用户352相关联。因此,可同时监控多个账户350和/或资源消耗用户352的资源使用。例如,公司实体可以具有与其雇员(例如,资源消耗用户352)相关联的账户350。因此,当雇员访问计算资源108时,使用与对应于公司实体的账户350相关联。通过将使用与账户350相关联,单个度量/属性收集器303可用于监控多个账户350,与将度量/属性收集器303与每个账户350和/或资源消耗用户352相关联相比,这可减少所消耗的资源。

[0063] 账户350可以与资源消耗用户352可访问的一个或更多个(例如,虚拟或物理的)机器354相关联,并且使用计算资源108操作。例如,机器354可包括操作系统112和/或应用114中的一个或更多个。例如,机器354可被配置为在虚拟化或物理环境中执行基于LINUX的操作系统、基于WINDOWS的操作系统、基于MAC的操作系统和/或任何其他操作系统。在一些情况下,机器354可被配置为在不执行客户操作系统的情况下执行一个或更多个工作负载110(例如,(多个)应用114)。例如,机器354可被配置为直接在物理机器上执行一个或更多个应用114而无任何形式的虚拟化或被配置为执行容器,一个或更多个应用114可被配置为在容器内执行。因此,容器通常可被描述为被配置以在机器354的一个或更多个内核上执行的隔离环境。容器可以在物理或虚拟机器上执行。

[0064] 机器集合管理器348允许多个机器354与至少一个机器集合356相关联。机器集合管理器348还可以允许多个容器或唯一标识的(例如,加标签的)工作负载与至少一个集合相关联。在一些情况下,多个机器集合356可与相应账户350相关联。例如,相应账户350的所有者(或授权用户)可将一个或更多个机器354添加到相应机器集合356。每个机器集合356可包括多个机器354,每个机器354被配置为执行例如不同的操作系统112和/或应用114(例如,可使用容器来执行应用114)。

[0065] 插件数据库338可以被配置成由度量/属性收集器303访问并且可以包括一个或更多个插件358(例如,插件358可以是驱动器、监听器、发布器等),该一个或更多个插件358允许度量/属性收集器303与一个或更多个计算系统104进行通信,使得资源使用例如可以被测量。每个计算系统104可被配置成包括一个或更多个机器354,每个机器354被配置为执行不同的操作系统112和/或应用114。换句话说,插件358允许度量/属性收集器303同时监控多个不同计算环境中的资源使用。例如,插件358可以包括LINUX插件、WINDOWS插件、MAC插件、SOLARIS插件、VMWARE插件、HYPER-V插件、CITRIX XEN插件、OPENSTACK插件、OPENSHIFT插件、KUBERNETES插件、DOCKER插件、PIVOTAL插件、AMAZON插件,MICROSOFT插件、GOOGLE插件、裸金属插件(例如,被配置成与非虚拟化机器直接通信的插件),和/或允许跨多个操作系统/环境来监控资源使用的任何其他插件。插件358还可以允许同时监控一个或更多个虚

拟化和/或物理环境。插件358还可以使得能够监控其他用户账户信息,其他用户账户信息包括计费信息、软件程序利用率、租用的软件许可证和/或任何其他附加用户账户信息。

[0066] 插件358可以被配置成被激活(例如,度量/属性收集器303能够使用相应的插件358与相应的计算系统104通信)或被停用(例如,度量/属性收集器303不能使用相应的插件358与相应的计算系统104通信)。

[0067] 机器管理器342可以允许将机器354分配和/或关联到相应的机器集合356。机器管理器342可分析(多个)机器集合356中的机器354以确定例如机器354的规范和/或操作系统。规范可包括例如与计算资源108相关的信息(例如,可用存储器、储存器等)。

[0068] 机器管理器342还可以搜索将由度量/属性收集器303监控的新机器。机器管理器342可以通过搜索与特定账户350和/或资源消耗用户352相关联的新IP地址来检测新机器。

[0069] 收集器数据库管理器344管理由度量/属性收集器303接收和/或生成的数据。例如,收集器数据库管理器344可以接收和/或使得储存与机器354、机器集合356、账户350和与机器354、机器集合356和账户350相关联的资源使用(例如,度量/属性)有关的一个或更多个记录(例如,数据)。收集器数据库管理器344还可删除数据。例如,收集器数据库管理器344可以删除与相应机器354、相应机器集合356和/或相应账户350相关联的记录。在一些情况下,收集器数据库管理器344可以截短与相应机器354、相应机器集合356和/或相应账户350相关联的记录。删除和/或截短可在例如一段时间已过去之后和/或在相应账户350的停用/终止之后发生。

[0070] 收集器数据库管理器344可以经由通信链路366与收集器数据库364通信并且使得一个或更多个记录被储存在收集器数据库364上。如图所示,收集器数据库364可以远离度量/属性收集器303(例如,在不同的服务器和/或不同的网络上操作)。然而,在一些情况下,收集器数据库364可以在度量/属性收集器303本地(例如,在同一服务器上、在同一网络上和/或在同一设备上操作)。经由通信链路366传输的数据可被加密以防止在传输期间对数据的未授权访问。在一些情况下,储存在收集器数据库364上的记录可以被加密以防止对记录的未授权访问。例如,可以使用透明数据加密(TDE)来加密储存在收集器数据库364上的记录。

[0071] 收集器同步器336可以与例如收集器数据库管理器344通信,并且可以复制和/或备份由收集器数据库管理器344接收的数据。例如,收集器同步器336可以将储存在收集器数据库364上和/或将要储存在收集器数据库364上的一个或更多个记录复制到收集器备份368上。在一些情况下,可以对收集器备份368上储存的记录进行加密以防止未授权的访问。当对一个或更多个记录做出改变时,收集器同步器336可以仅复制和/或备份对收集器备份368的改变。结果,收集器同步器336可以不将整个记录复制到收集器备份368。如图所示,收集器同步器336可以经由通信链路370与收集器备份368通信。经由通信链路370传输的数据可被加密以防止在传输期间对数据的未授权访问。如图所示,收集器备份368可以远离度量/属性收集器303(例如,在不同的服务器上、在不同的网络中、和/或在不同的设备上操作)。然而,在一些情况下,收集器备份368可在度量/属性收集器303本地(例如,在同一服务器上、在同一网络上和/或在同一设备上操作)。

[0072] 在一些情况下,可设置至少一个备份收集器397。备份收集器397可以与度量/属性收集器303基本相同地操作,并且可以被配置为与度量/属性收集器303并行(或同时)运行。

在一些情况下,备份收集器397被配置成远离度量/属性收集器303(例如,在不同的服务器上、在不同的网络中、和/或在不同的设备上操作)。因此,在度量/属性收集器303故障的情况下,计量器118可以继续监控资源使用。换句话说,备份收集器397可通常被描述为提供冗余。在一些情况下,可存在两个或更多个备份收集器397,使得可实现例如n+2冗余。因此,备份收集器397可包括本文关于度量/属性收集器303所讨论的特征中的任何一个或更多个。备份收集器397可经由通信链路399与计算系统104通信。经由通信链路399传输的数据可被加密以防止在传输期间对数据的未授权访问。

[0073] 备份收集器397还可以被配置为生成独立记录。可以在例如审计过程期间和/或在数据恢复过程中(例如,在度量/属性收集器303经历故障的情况下)使用独立的记录。在一些情况下,收集器备份368可以包括备份收集器397。

[0074] 如图所示,度量/属性收集器303还可以包括收集器自动器340。收集器自动器340可以使度量/属性收集器303的一个或更多个特征自动化。例如,收集器自动器340可以被配置成接收使度量/属性收集器303执行预定动作(例如,开始或终止接收度量/属性、开始数据库清理、启用或禁用对计算资源108的访问、激活或停用一个或更多个插件358和/或任何其他动作)的触发事件。触发事件可以包括例如发现(多个)新机器、合同条款期满、使用预定量的资源、将新机器添加到相应的机器集合356等。在一些情况下,触发事件可以是一段时间的期满。例如,每1、2、3、4或5秒(或任何其他合适的时间间隔),度量/属性收集器303可以搜索对应于新机器的新的互联网协议(IP)地址(和/或任何其他标识符)。在这些情况下,度量/属性收集器303通常可以被描述为自更新。

[0075] 如图还示出的,管理员用户接口372可通信地耦接至收集器核332。管理员用户接口372可以为度量/属性收集器303的操作者提供接口,以控制和/或修改度量/属性收集器303和/或查看由度量/属性收集器303生成和/或接收的数据。例如,管理员用户接口372可被配置成基于每账户350或资源消耗用户352来显示和/或使得生成使用摘要。管理员用户接口372还可用于例如手动地添加一个或更多个机器354,调整和/或删除一个或更多个记录、添加或移除新账户350、添加或移除插件358、调整/创建用于收集器自动器340的触发器、响应于例如合同的条款和/或任何其他管理任务而在预定时间继续和/或中止监控资源使用。

[0076] 如图所示,管理员用户接口372可经由通信链路374与收集器核332通信。经由通信链路374传输的数据可被加密以防止在传输期间对数据的未授权访问。如图所示,管理员用户接口372可远离度量/属性收集器303(例如,在不同的服务器上、在不同的网络中、和/或在不同的设备上操作)。然而,在一些情况下,管理员用户接口372可在度量/属性收集器303本地(例如,在同一服务器上、在同一网络上和/或在同一设备上操作)。

[0077] 在一些情况下,例如,可以在网络浏览器上生成和显示管理用户接口372。作为另一示例,在一些情况下,管理用户接口372可以是在计算机上执行的通过网络连接(例如,互联网)与度量/属性收集器303通信的应用。作为又一示例,在一些情况下,管理用户接口372可以是在本地(例如,在与度量/属性收集器303相同的设备上)执行的应用。

[0078] 图3C示出了分析平台360的示例,该分析平台360可以是图3A的分析平台312的示例。分析平台360通常可以被描述为被配置为将在一段时间内接收的多个度量/属性归一化(例如,转换)为归一化单元,并且对归一化的度量/属性求和以生成表示为该时间段的单个

归一化单元的总资源使用值。换句话说,分析平台360可被配置成生成表示在一时间段上的总资源使用的单个使用值。分析平台360可包括硬件、软件和/或固件的任何组合。在一些情况下,分析平台360可在与度量/属性收集器302相同的服务器和/或相同的网络上执行。在其他情况下,分析平台360可在与度量/属性收集器302不同的服务器和/或不同的网络上(例如,在用户的网络和/或服务器上)执行。

[0079] 分析平台360可被配置为从度量/属性收集器302和/或一个或更多个备份收集器397中的一个或更多个接收表示计算资源108的使用的一个或更多个度量/属性。在一些情况下,分析平台360可从计算系统104接收一个或更多个度量/属性。在这些情况下,计量器118可以不包括度量/属性收集器302,并且分析平台360可以包括插件数据库338。分析平台360可接收呈例如基于JavaScript语言的轻量级的数据交换格式(JSON)文件、逗号分隔值(CSV)文件和/或任何其他文件类型的形式的度量/属性。

[0080] 分析平台360可以被配置成与计算系统104、设备102和/或通信地耦接到网络106的任何其他设备中的任何一个或更多个包括在一起。例如,设备102可以包括分析平台360。当设备102包括分析平台360时,与不在设备102上执行分析平台360时相比,分析平台360的用户能够通过对用户的网络安全性的更少修改来使用计量器118。在一些情况下,当设备102包括分析平台360时,分析平台360可被配置成分析例如局域网上的一个或更多个用户的资源使用,其中用户中的每一个都具有例如相应的设备。另外或替代地,当设备102包括分析平台360和度量/属性收集器302两者时,仅表示计算资源108的使用的度量/属性可能需要通过公共网络(例如,互联网)传输。

[0081] 分析平台360可包括至少一个分析核376、至少一个分析度量接收器378、至少一个分析同步器380、至少一个分析数据库管理器382、至少一个分析自动器384、至少一个使用分析器386、至少一个通知生成器388、至少一个订阅管理器390以及至少一个交换/合同管理器393。分析核376通常可以被描述为与分析度量接收器378、分析同步器380、分析数据库管理器382、分析自动器384、使用分析器386、通知生成器388和订阅管理器390中的每一个进行通信和/或管理分析度量接收器378、分析同步器380、分析数据库管理器382、分析自动器384、使用分析器386、通知生成器388和订阅管理器390中的每一个。

[0082] 如图所示,分析核376通信地耦接至分析度量接收器378。分析度量接收器378从度量/属性收集器302接收表示计算资源108的使用的度量/属性。分析度量接收器378可包括计算器305中的一个或更多个。在一些情况下,分析度量接收器378可被配置以生成表示为单个归一化单元的总资源使用值。在这些情况下,分析核376可被配置以接收表示为单个归一化单元的总资源使用值。在其他情况下,分析核376可被配置为生成表示为单个归一化单元的总资源使用值。

[0083] 分析核376和/或分析度量接收器378可以与使用分析器386通信。使用分析器386监控与相应账户350和/或资源消耗用户352相关联的使用,以开发(或确定)正常使用、高峰使用、和/或低谷使用。换句话说,正常使用、高峰使用和/或低谷使用可至少部分地基于对应于计算资源108中的相应计算资源所接收的度量/属性。在一些情况下,例如,高峰使用可表示在一时间段(例如,一小时、一天、一周、一个月和/或任何其他合适的时间段)内的最大使用,低谷使用可表示一段时间(例如,一小时、一天、一周、一个月和/或任何其他合适的时间段)内的最小使用,以及正常使用可以表示一段时间(例如,一小时、一天、一周、一个月

和/或任何其他合适的时间段)内的平均使用。在一些情况下,高峰使用可以近似一段时间内的最大使用(例如,是实际最大使用的5%、10%或15%内的使用值),低谷使用可以近似一段时间内的最小使用(例如,是实际最小使用的5%、10%或15%内的使用值)。

[0084] 正常使用、高峰使用和/或低谷使用可对应于单个计算资源、多个计算资源和/或构成计算资源的一个或更多个子资源(例如,GPU210的图形处理器和图形存储器中的一个或更多个)。例如,可针对多个计算资源中的每一个(例如,CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个中的每一个)生成正常使用、高峰使用和/或低谷使用。在这些情况下,每个资源的正常使用、高峰使用和/或低谷使用可表示为归一化单元和/或(多个)非归一化单元(例如,表示为对应于各个资源的单元)。作为另一示例,可生成对应于多个计算资源(例如,CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个中的两个或更多个)的正常使用、高峰使用和/或低谷使用。在这些情况下,正常使用、高峰使用和/或低谷使用可表示为单个归一化单元。

[0085] 在一些情况下,使用分析器386可被配置为生成总高峰使用平均值。总高峰使用平均值一般可被描述为对应于两个或更多个相应的预定时间段的两个或更多个高峰使用值的平均值。例如,可针对一周中的每一天确定高峰使用且对高峰使用求平均以确定一周的总高峰使用平均值。另外或替代地,使用分析器386可被配置以生成总低谷使用平均值。总低谷使用平均值通常可被描述为对应于两个或更多个相应的预定时间段的两个或更多个低谷使用值的平均值。例如,可为一周中的每天确定低谷使用,并对低谷使用求平均以确定一周的总低谷使用平均值。

[0086] 总高峰使用平均值和/或总低谷使用平均值可对应于各个计算资源和/或多个计算资源。例如,可针对多个计算资源中的每一个(例如,CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个中的每一个)生成总高峰使用平均值和/或总低谷使用平均值。在这些情况下,每个资源的总高峰使用平均值和/或总低谷使用平均值可表示为归一化单元和/或(多个)非归一化单元(例如,表示为对应于各个资源的单元)。作为另一示例,可生成对应于多个计算资源(例如,CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个中的两个或更多个)的总高峰使用平均值和/或总低谷使用平均值。在这些情况下,总高峰使用平均值和/或总低谷使用平均值可以表示为单个归一化单元。

[0087] 正常平均使用通常可被描述为平均资源使用,该平均资源使用通过在重复的时间间隔内的时间段内对实际资源使用进行多次测量并对测量值求平均来计算。例如,可在每天上午11:00测量实际资源使用达5天,并且可基于这些测量值来计算上午11:00的平均使用。高峰平均使用通常可以被描述为平均资源使用,该平均资源使用是通过在重复的时间间隔内的时间段内对实际资源使用进行多次测量,选择具有最高使用(例如,资源使用的最高5%、10%、15%、20%、25%、35%、40%、50%或60%)的测量值的部分,并计算所选部分的平均值来计算的。例如,可在每天上午11:00测量实际资源使用达五天,且可基于三个最高测量值来计算上午11:00的高峰平均使用。低谷平均使用通常可以被描述为平均资源使用,该平均资源使用是通过在重复的时间间隔内的时间段内对实际资源使用进行多次测量,选择具有最低使用(例如,资源使用的最低5%、10%、15%、20%、25%、35%、40%、50%

或60％)的测量值的部分,并计算所选部分的平均值来计算的。例如,可在每天上午11:00测量实际资源使用达五天,且可基于三个最低测量值来计算上午11:00的低谷平均使用。

[0088] 正常平均使用、高峰平均使用和/或低谷平均使用可对应于单个计算资源和/或多个计算资源。例如,可针对多个计算资源中的每一个(例如,CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个中的每一个)生成正常平均使用、高峰平均使用和/或低谷平均使用。在这些情况下,每个资源的正常平均使用、高峰平均使用和/或低谷平均使用可表示为归一化单元和/或(多个)非归一化单元(例如,表示为对应于各个资源的单元)。作为另一示例,可生成对应于多个计算资源(例如,CPU 202、存储器204、储存器206、网络接口208、GPU 210和/或其他资源209中的一个或更多个中的两个或更多个)的正常平均使用、高峰平均使用和/或低谷平均使用。在这些情况下,正常平均使用、高峰平均使用和/或低谷平均使用可表示为单个归一化单元。

[0089] 使用分析器386可分别基于正常平均使用、高峰平均使用和低谷平均使用来生成正常平均使用模式392、高峰平均使用模式394和低谷平均使用模式395(为了清楚起见,在图3D中以图形方式示出)。正常平均使用模式392可以是包括与不同预定时间(或预定时间段)对应的多个正常平均使用值的数据集。高峰平均使用模式394可以是包括对应于不同预定时间(或预定时间段)的多个高峰平均使用值的数据集。低谷平均使用模式395可以是包括对应于不同的预定时间(或预定时间段)的多个低谷平均使用值的数据集。

[0090] 可将对应于正常平均使用模式392的数据集与对应于高峰平均使用模式394的数据集进行比较,以确定一个或更多个预定时间(或预定时间段)的阈值使用值或范围。例如,可通过取在相应的预定时间的正常平均使用值和高峰平均使用值的平均值来获得预定时间的阈值使用值。作为另一示例,阈值使用范围可以被描述为在相应的预定时间从正常平均使用值延伸到高峰平均使用值的范围。

[0091] 另外或替代地,可将对应于低谷平均使用模式395的数据集与对应于高峰平均使用模式394的数据集进行比较,以确定一个或更多个预定时间(或预定时间段)的阈值使用值或范围。例如,可通过取相应的预定时间的低谷平均使用值和高峰平均使用值的平均值来获得预定时间的阈值使用值。作为另一示例,阈值使用范围可以被描述为在相应的预定时间从低谷平均使用值延伸到高峰平均使用值的范围。

[0092] 另外地或替代地,可将对应于正常平均使用模式392的数据集与对应于低谷平均使用模式395的数据集进行比较,以确定一个或更多个预定时间(或预定时间段)的阈值使用值或范围。例如,可通过取相应的预定时间的正常平均使用值和低谷平均使用值的平均值来获得预定时间的阈值使用值。作为另一示例,阈值使用范围可被描述为在相应的预定时间从低谷平均使用值延伸到正常平均使用值的范围。

[0093] 另外地或替代地,阈值使用值可以是低谷平均使用值、高峰平均使用值、低谷使用值(即,在预定时间内的最小使用)或总低谷使用平均值,和/或高峰使用值(即,在预定时间内的最大使用)或总高峰使用平均值。

[0094] 通知生成器388可被配置为响应于例如在预定时间(或预定时间间隔)的资源使用满足、超过和/或未能超过阈值使用值和/或资源使用落在阈值使用范围之外的一个或更多个而使通知生成。换句话说,可使得至少部分地基于资源使用与阈值或阈值使用范围的比较来生成通知。例如,检测超过阈值使用值的使用可指示未授权的第三方使用资源,并且检

测不超过低谷平均使用值的使用可指示硬件故障(例如,计算资源108的至少一部分可能出故障)。

[0095] 在一些情况下,使用分析器386可被配置为识别异常值(例如,落在阈值使用范围之外的资源使用)。例如,使用分析器386可利用机器学习来识别高使用和低使用的时间段。在高使用和/或低使用的时间段中,如果使用持续至少一段时间(例如,10分钟、30分钟、60分钟、2小时、和/或任何其他时间跨度)不超过高峰平均使用值和/或不超过低谷平均使用值,则使用分析器386可以将使用识别为异常值。因此,通知生成器388可被配置为响应于异常值的检测而不生成通知。此外,当识别出异常值时,使用分析器386可以从正常平均使用模式392、高峰平均使用模式394和/或低谷平均使用模式395的计算中排除异常值,除非例如多次检测到异常值。因此,正常平均使用模式392、高峰平均使用模式394和/或低谷平均使用模式395均可通常被描述为使用机器学习来生成。

[0096] 在一些情况下,使用分析器386可以分析计量器118的其他用户的使用以确定其他用户是否正在经历类似的使用模式。例如,如果多个用户正经历不超过低谷平均资源值的资源使用,那么这可指示硬件故障(例如,计算资源108的至少一部分可能出故障)。在这种情况下,通知生成器388可使通知生成。

[0097] 所生成的通知可以是例如,经由例如电话呼叫、SMS文本消息、电子邮件、传真、触感(例如,经由智能手表、健身追踪器、远程健康监控设备和/或任何其他可穿戴设备)传输至账户350中的一个或更多个的所有者/用户的警报的形式,在应用或软件内生成的用于管理分析平台360和/或度量/属性收集器302的警报(例如,应用内的消息),和/或任何其他类型的警报。

[0098] 使用分析器386可使用例如机器学习来确定在一段时间内定期发生的高资源使用的时段和低资源使用的时段。当识别出高资源使用的时段和低资源使用的时段时,可以调整例如使得生成警报的阈值使用。例如,在高资源使用的已知时段期间,用于生成警报的阈值使用可被调整为高于在低资源使用的已知时段期间。

[0099] 使用分析器386还可以被配置为生成基准(例如,平均成本指数),所述基准表示针对一个或更多个用户和/或账户所消耗的资源的每单元成本。使用分析器386可以接收被表示为与一个或更多个账户350、资源消耗用户352和/或机器集合356相关联的单个归一化单元的总资源使用值。在一些情况下,使用分析器386可以接收对应于工作负载类型(例如,SPARK集群工作负载、ELASTICSEARCH工作负载、数据仓储工作负载、灾难恢复工作负载、开发者工作负载、质量保证工作负载、生产工作负载等)、服务提供商(例如,AMAZON、MICROSOFT、GOOGLE和/或任何其他服务提供商)、硬件提供商(例如,DELL、SUPERMICRO、HP和/或任何其他硬件提供商)、硬件配置、管理程序(例如,VMWARE、CITRIX、OPENSTRACK和/或任何其他管理程序)和/或组(例如,诸如工程、销售或营销部门的组织内的业务单元)的总资源使用值。

[0100] 使用分析器386还可以接收与正在消耗的资源相关联的成本数据。可以从计算系统104(例如,服务提供商)、从用户输入和/或任何其他源接收成本数据。成本数据可以表示与相应账户350、资源消耗用户352、机器集合356、工作负载类型和/或组相关联的总资源使用的成本。与正在被消耗的资源相关联的成本数据可包含多个托管成本、硬件成本、联网成本、云代管成本、劳动力成本、软件许可证成本、合规性成本、管理的服务成本,咨询者成本、

支持成本、开发者/操作教育成本、网络经济成本、迁移机会成本和/或被消耗的资源的任何其他相关联的成本中的一个或更多个。使用分析器386然后可以将总资源使用的成本除以总资源使用值以获得每单元消耗值的成本。使用分析器386可利用每单元消耗的成本来生成基准(例如,平均成本指数)。

[0101] 然后可以将每单元消耗值的成本与例如市场基准(例如,跨一个或更多个计量器118的所有用户的每单元成本的平均值)、组或行业基准(例如,跨预定行业或组内的所有用户的每单元成本的平均值)、工作负载基准(例如,跨执行预定工作负载的所有用户的每单元成本的平均值)、服务提供商(例如,AMAZON、MICROSOFT、GOOGLE和/或任何其他服务提供商)基准和/或类似基准进行比较。组或行业基准可基于例如公司大小(例如,市值(market cap)、年收入等)、特定公司内或跨多个公司(例如,销售部门、工程部门等)的业务单元或部门和/或任何其他组。

[0102] 在一些情况下,使用分析器386还可以被配置为生成市场基准、组或行业基准、服务提供商基准和/或工作负载基准。例如,使用分析器386可被配置为访问与对应于一个或更多个不同用户的一个或更多个账户350的资源使用和资源成本相关的数据。

[0103] 在一些情况下,使用分析器386可以与分析平台360分开。例如,使用分析器386可被配置为与多个分析平台360和/或度量/属性收集器302通信。在一些情况下,可以存在市场使用分析器,该市场使用分析器远离分析平台360并且被配置为确定跨多个不同用户的特定组或行业、特定账户、特定工作负载等的平均使用和平均成本。在这些情况下,还可以存在本地使用分析器,该本地使用分析器被配置为确定特定分析平台360的平均使用和平均成本。

[0104] 在一些情况下,使用分析器386还可被配置为分析正在被消耗的计算资源108以确定正在计算资源108上执行的工作负载110是否正在有效地利用计算资源108。例如,使用分析器386可分析计算资源108的规范(例如,处理器速度和/或大小、存储器速度和/或大小、GPU速度和/或大小、储存器速度和/或大小等)且将这些规范与工作负载110的要求进行比较,使得使用分析器386可生成优化的计算资源配置(例如,最佳估计配置)。在此情况下,使用分析器386可被配置为使得生成待发送到用户的通知,其中该通知可包含优化的计算资源配置的描述。在一些情况下,使用分析器386可被配置为自动地将工作负载110转换到优化的计算资源配置。作为另一示例,计算资源108可包括专用图形处理单元,然而,在计算资源108上执行的应用114和/或操作系统112可能不需要专用图形处理单元来高效地操作。在这种情况下,使用分析器386可以被配置为使得生成将要发送给用户的通知。在一些情况下,使用分析器386可被配置为将工作负载110自动移动到不包括专用图形处理单元的另一计算系统104。因此,所消耗的计算资源的每单元的成本可以被降低。

[0105] 在一些情况下,使用分析器386可进一步被配置为基于待执行的工作负载110生成计算资源配置推荐。计算资源配置推荐可与计算资源消耗值的每单元的成本和/或相应工作负载110的估计性能等级相关联。因此,用户可以比较多个资源供应商之间和/或可从特定资源供应商获得的多个计算机资源选项之间的资源成本和/或性能。

[0106] 在一些情况下,当用户正在预先购买预定量的资源时,使用分析器386可被配置为分析用户的资源使用。在这些情况下,使用分析器386可以将预先购买的资源的量转换成表示可用资源的最大量的归一化值,并且将预先购买的资源的归一化值与高峰使用值进行比

较。换句话说,使用分析器386可被配置为在用户预先购买预定量的资源时确定资源供应是否足够满足用户的高峰资源需求。因此,基于该比较,使用分析器386可被配置为提供关于要预先购买的资源量的推荐。

[0107] 订阅管理器390管理和维护与计量器118的用户有关的订阅信息。例如,订阅可确定多少资源消耗用户352可被分配给相应账户350,确定多少账户可以被分配给相应用户(例如,公司实体),确定多少机器354可由相应用户访问,确定多少机器354可被分配给相应机器集合356,确定相应用户可具有多少机器合集356,将登录信息与相应用户相关联,确定登录信息的有效性(例如,登录信息是否已经期满),和/或与订阅管理相关联的任何其他任务。

[0108] 交换/合同管理器393被配置为协调两方(例如,买方和卖方)之间的合同。交换/合同管理器393可以被配置为根据合同的条款禁用计量器118。当计量器118已经被禁用时,可以禁用用户对计算资源108的访问。交换/合同管理器393还可以被配置成审核资源使用。资源使用的审核可涉及确认被指示为正在被使用的资源实际上被使用和/或确认使用是根据合同的。审核过程可涉及将由度量/属性收集器302收集的数据与由第二收集器(例如,备份收集器)生成的数据进行比较。

[0109] 分析自动器384可以使分析平台360的一个或更多个特征自动化。例如,分析自动器384可以被配置为接收使分析平台360执行预定动作的触发事件。预定动作可包括例如启用或禁用对计算资源108的访问,激活或停用度量/属性收集器302(例如,度量接收器334)的至少一部分,激活或停用一个或更多个插件358(例如,被配置为使得能够与计算系统104进行通信的一个或更多个插件358),开始数据库清理和/或任何其他动作中的一个或更多个。触发事件可以包括例如接收新合同、接收新订阅、来自管理员用户接口372的用户输入、预定时间(例如,新合同/订阅的开始和/或任何其他预定时间)、召回请求等。在一些情况下,触发事件可以是一段时间的期满(例如,合同条款的终止)。例如,分析平台360可响应于合同条款终止而执行预定动作。

[0110] 如图所示,分析数据库管理器382与分析核376和分析同步器380通信。分析数据库管理器382管理由分析平台360接收和/或生成的数据。例如,分析数据库管理器可以管理和/或使得储存与订阅、合同、度量/属性、使用模式(例如,正常平均使用模式392、高峰平均使用模式394或低谷平均使用模式395)、通知阈值、触发事件和/或由分析平台360接收或生成的任何其他数据相关的一个或更多个记录(例如,数据)。分析数据库管理器382还可以删除和/或截短数据。例如,分析数据库管理器382可以删除和/或截短与订阅、合同、度量/属性、使用模式(例如,正常平均使用模式392、高峰平均使用模式394或低谷平均使用模式395)、通知阈值、和/或由分析平台360接收的任何其他数据有关的记录。删除和/或截短可以在例如一段时间已经过去之后和/或在例如订阅或合同的停用/终止之后发生。

[0111] 分析数据库管理器382可以经由通信链路398与分析数据库396通信并且使得一个或更多个记录被储存在分析数据库396上。如图所示,分析数据库396可以远离分析平台360(例如,在不同的服务器上、在不同的网络中、和/或在不同的设备上操作)。然而,在一些情况下,分析数据库396可在分析平台360本地(例如,在同一服务器上、在同一网络上和/或在同一设备上操作)。经由通信链路398传输的数据可被加密以防止传输期间的未授权访问。在一些情况下,可以对储存在分析数据库396上的记录进行加密以防止对这些记录的未授

权访问。例如,可以使用透明数据加密(TDE)来加密储存在分析数据库396上的记录。

[0112] 分析同步器380可以与分析度量接收器378和/或分析数据库管理器382进行通信,并且可以复制和/或备份由分析平台360接收的数据(例如,由分析数据库管理器382接收的数据)。例如,分析同步器380可以将储存在和/或将被储存在分析数据库396上的记录中的一个或更多个复制和/或备份到分析备份381上。在一些情况下,可以对储存在分析备份381上的记录进行加密以防止未授权访问。当对一个或更多个记录进行改变时,分析同步器380可以仅将改变备份到分析备份381。如图所示,分析同步器380可以经由通信链路383与分析备份381通信。经由通信链路383传输的数据可被加密以防止在传输期间对数据的未授权访问。如图所示,分析备份381可远离分析平台360(例如,在不同的服务器上、在不同的网络中、和/或在不同的设备上操作)。然而,在一些情况下,分析备份381可以在分析平台360本地(例如,在同一服务器上、在同一网络上和/或在同一设备上操作)。

[0113] 在一些情况下,可设置与分析平台360基本相同地操作的备份分析平台355。因此,可维持独立的记录。独立的记录可以例如在审计过程中和/或在数据恢复过程中(例如,在分析平台360经历故障的情况下)使用。备份分析平台355可以使用通信链路353通信地耦接到度量/属性收集器302。

[0114] 在一些情况下,分析备份381还可包括与分析平台360基本相同地操作的备份分析平台。换句话说,分析备份381可以包括独立的分析平台。因此,可维持独立的记录。独立的记录可以例如在审计过程中和/或在数据恢复过程中

[0115] (例如,在分析平台360经历故障的情况下)使用。

[0116] 如图所示,使用预测器385可以经由通信链路387通信地耦接至分析核376。使用预测器385可以估计相应用户(例如,拥有或控制一个或更多个账户350的用户)、相应账户350和/或相应资源消耗用户352的未来使用。使用预测器385可以使用例如正常平均使用模式392、高峰平均使用模式394和/或低谷平均使用模式395来预测未来使用。附加地或可替代地,使用预测器385可利用机器学习来分析历史资源使用以预测未来使用。例如,使用预测器385可以分析用于模式和/或激增(例如,资源使用的增加)的使用。一旦预测未来的使用,使用预测器385还可以估计与预测的使用相对应的预测的经济成本。

[0117] 如图所示,分析用户接口389可经由通信链路391与分析核376通信。分析用户接口389可被配置为生成呈现与分析平台360和/或度量/属性收集器302有关的信息的显示。例如,分析用户接口389可被配置为显示使用预测,正常平均使用模式392(例如,作为曲线图或绘图)、高峰平均使用模式394(例如,作为曲线图或绘图),低谷平均使用模式395(例如,作为曲线图或绘图)、当前合同条款,当前订阅、活动和/或非活动的通知/提醒、和/或关于分析平台360的任何其他数据。在一些情况下,分析用户接口389可被配置成允许用户修改分析平台360的各种特征。例如,分析用户接口389可以被配置为使得用户可以改变和/或取消一个或更多个订阅、改变和/或取消一个或更多个合同、改变、添加和/或移除通知阈值和/或分析自动器384触发和/或改变分析平台360的任何其他特征。

[0118] 如图所示,分析用户接口389可以远离分析平台360(例如,在不同的服务器上、在不同的网络中、和/或在不同的设备上操作)。然而,在一些情况下,分析用户接口389可以在分析平台360本地(例如,在同一服务器上、在同一网络上和/或在同一设备上操作)。通过通信链路391传输的数据可被加密以防止对数据的未授权访问。

[0119] 例如,在一些情况下,分析用户接口389可被生成并显示在网络浏览器上。作为另一示例,在一些情况下,分析用户接口389可以是在通过网络连接(例如,互联网)与分析平台360通信的计算机上执行的应用。作为又一示例,在一些情况下,分析用户接口389可以是在本地(例如,在与分析平台360相同的硬件上)执行的应用。

[0120] 通信链路330、353、362、366、370、374、383、387、391、398和/或399可以在建立通信时利用安全套接层(SSL)安全协议。如本文所讨论的,通信链路330、353、362、366、370、374、383、387、391、398和/或399可以各自携带经加密的数据。因此,通信链路330、353、362、366、370、374、383、387、391、398和/或399中的一个或更多个通常可以被描述为安全通信链路(例如,通信链路330、353、362、366、370、374、383、387、391、398和/或399使用SSL安全协议中的至少一个和/或所传输的数据被加密)。

[0121] 度量/属性收集器302和分析平台360可以在软件、固件、硬件和/或其组合中实现。例如,度量/属性收集器302和分析平台360可以被实现为软件,该软件被配置为储存在一个或更多个存储器(例如,任何类型的有形的、非暂时性的储存介质,并且可以包括磁记录介质(例如,硬盘驱动器)、光盘、诸如只读存储器(ROM)的半导体设备、诸如动态和静态RAM的随机存取存储器(RAM)、可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、闪存、磁卡或光卡,或用于储存电子指令的任何类型的储存媒介中的任何一个或更多个)上并且由一个或更多个处理器(例如,由英特尔、高级微设备公司、IBM、ARM、甲骨文提供销售的处理器和/或任何其他处理器)执行以执行一个或更多个操作。作为另一示例,度量/属性收集器302和分析平台360可被实现为电路(例如,专用集成电路)。度量/属性收集器302和分析平台360可在相同或不同的机器、服务器和/或网络上实现。

[0122] 图4示出了被配置为生成如关于图3A所讨论的转换单元的转换单元生成器400的示意性示例。转换单元生成器400可以与计算系统104、设备102以及网络106分离。例如,转换单元生成器400可以是第三方系统的一部分,使得转换单元可被生成并作为固定值输入到分析平台312中。如图所示,转换单元生成器400可至少包括第一资源转换单元生成器402和第二资源转换单元生成器404。第一资源转换单元生成器402可生成例如CPU转换单元314、存储器转换单元316、储存器转换单元318、网络转换单元320和/或磁盘IO转换单元328中的一个或更多个。第二资源转换单元生成器404可以生成例如GPU处理器转换单元322、GPU存储器转换单元324和/或GPU存储器速度转换单元326中的一个或更多个。

[0123] 如图所示,第一资源转换单元生成器402包括CPU转换单元生成器406、存储器转换单元生成器408、储存器转换单元生成器410、网络转换单元生成器412和磁盘IO转换单元生成器414。

[0124] CPU转换单元生成器406可以至少部分地基于CPU分配416和CPU计算部分418来生成CPU转换单元314。例如,CPU转换单元314可以通过由CPU计算部分418划分CPU分配416来生成。

[0125] CPU分配416对应于用户可访问的CPU 202的时钟速度。例如,可以仅将物理处理器的一部分分配给一个或更多个用户的工作负载。换句话说,被分配的可用时钟速度可测量小于物理处理器的默认时钟速度。CPU计算部分418通常可以对应于表示由工作负载110中的一个或更多个使用的CPU资源的权重。在一些情况下,CPU计算部分418可以是实验推导的值,该实验推导的值是基于对若干用户的工作负载的分析的。照此,在一些情况下,CPU计算

部分418通常可以被描述为在预定时间是可调整的以应对工作负载的变化的动态权重。

[0126] 存储器转换单元生成器408可以至少部分地基于存储器分配420和存储器计算部分422来生成存储器转换单元316。例如,可以通过由存储器计算部分422划分存储器分配420来生成存储器转换单元316。

[0127] 存储器分配420对应于用户可访问的存储器204的量。例如,仅物理存储器的一部分可以被分配给一个或更多个用户的工作负载。换句话说,被分配的存储器可测量小于可用的总物理存储器。存储器计算部分422通常可以对应于表示由工作负载110中的一个或更多个使用的存储器资源的权重。在一些情况下,存储器计算部分422可以是实验推导的值,该实验推导的值是基于对若干用户的工作负载的分析的。照此,在一些情况下,存储器计算部分422通常可以被描述为动态权重,该动态权重在预定时间是可调整的以应对工作负载的变化。

[0128] 储存器转换单元生成器410可以至少部分地基于储存器分配424和储存器计算部分426来生成储存器转换单元318。例如,可以通过由储存器计算部分426划分储存器分配424来生成储存器转换单元318。

[0129] 储存器分配424对应于用户可访问的储存器206的量。例如,仅物理储存器的一部分可以被分配给一个或更多个用户的工作负载。换句话说,被分配的储存器可以测量少于可用的总物理储存器。储存器计算部分426可通常对应于表示由工作负载110中的一个或更多个使用的储存器资源的权重。在一些情况下,储存器计算部分426可以是实验推导的值,该实验推导的值是基于对若干用户的工作负载的分析的。照此,在一些情况下,储存器计算部分426通常可以被描述为在预定时间是可调整的以应对工作负载的变化的动态权重。

[0130] 网络转换单元生成器412可至少部分地基于网络分配428和网络计算部分430来生成网络转换单元320。例如,可通过由网络计算部分430划分网络分配428来生成网络转换单元320。

[0131] 网络分配428对应于用户可访问的网络带宽。例如,仅总物理网络带宽的一部分可以被分配给一个或更多个用户的工作负载。换句话说,被分配的网络带宽可测量小于可用的总物理网络带宽。网络计算部分430通常可以对应于表示由工作负载110中的一个或更多个使用的网络资源(例如,带宽)的权重。在一些情况下,网络计算部分430可以是实验推导的值,该实验推导的值是基于对若干用户的工作负载的分析的。照此,在一些情况下,网络计算部分430通常可以被描述为动态权重,该动态权重在预定时间是可调整的以应对工作负载的变化。

[0132] 磁盘Io转换单元生成器414可以至少部分地基于磁盘IO分配432和磁盘IO计算部分434生成磁盘IO转换单元328。例如,可以通过由磁盘IO计算部434划分磁盘IO分配432来生成磁盘IO转换单元328。

[0133] 磁盘IO分配432对应于用户可访问的磁盘带宽(例如,对储存器206的读取和写入)。例如,仅总物理磁盘带宽的一部分可以被分配给一个或更多个用户的工作负载。换句话说,被分配的磁盘带宽可以测量小于可用的总物理磁盘带宽。磁盘IO计算部分434通常可以对应于表示由工作负载110中的一个或更多个使用的磁盘IO资源(例如,磁盘带宽)的权重。在一些情况下,磁盘计算部分434可以是实验推导的值,该实验推导的值是基于对若干用户的工作负载的分析的。照此,在一些情况下,磁盘计算部分434通常可以被描述为在预

定时间是可调整的以应对工作负载的变化的动态权重。

[0134] CPU计算部分418、存储器计算部分422、储存器计算部分426、网络计算部分430和磁盘IO计算部分434的和可以总计达100。换句话说,通过将计算部分除以100,每个计算部分可表示为由第一资源转换单元生成器402评估的总计算资源的百分比。

[0135] 如图所示,第二资源转换单元生成器404包括GPU处理器转换单元生成器436、GPU存储器转换单元生成器438和GPU存储器速度转换单元生成器440。

[0136] GPU处理器转换单元生成器436可以至少部分地基于GPU处理器分配442和GPU处理器计算部分444来生成GPU处理器转换单元322。例如,可通过由GPU处理器计算部分444划分GPU处理器分配442来生成GPU处理器转换单元322。

[0137] GPU处理器分配442对应于用户可访问的GPU 210的时钟速度。例如,仅物理GPU处理器的一部分可以被分配给一个或更多个用户的工作负载。换句话说,被分配的GPU处理器的可用时钟速度可测量小于物理GPU处理器的默认时钟速度。GPU处理器计算部分444通常可以对应于表示由工作负载110中的一个或更多个使用的GPU处理器资源的权重。在一些情况下,GPU计算部分444可以是实验推导的值,该实验推导的值是基于对若干用户的工作负载的分析的。照此,在一些情况下,GPU计算部分444通常可以被描述为动态权重,该动态权重在预定时间是可调整的以应对工作负载的变化。

[0138] GPU存储器转换单元生成器438可以至少部分地基于GPU存储器分配446和GPU存储器计算部分448来生成GPU存储器转换单元324。例如,可通过由GPU存储器计算部分448划分GPU存储器分配446来生成GPU存储器转换单元324。

[0139] GPU存储器分配446对应于用户可访问的GPU 210的存储器的量。例如,仅GPU 210的物理存储器的一部分可被分配给一个或更多个用户的工作负载。换句话说,被分配的GPU 210的存储器可测量小于可用的GPU210的总物理存储器。GPU存储器计算部分448通常可以对应于表示由工作负载110中的一个或更多个使用的存储器资源的权重。在一些情况下,GPU存储器计算部分448可以是实验推导的值,该实验推导的值是基于对若干用户的工作负载的分析的。照此,在一些情况下,GPU存储器计算部分448通常可以被描述为动态权重,该动态权重在预定时间是可调整的以应对工作负载的变化。

[0140] GPU存储器速度转换单元生成器440可以至少部分地基于GPU存储器速度分配450和GPU存储器速度计算部分452来生成GPU存储器速度转换单元326。例如,可以通过由GPU存储器速度计算部分452划分GPU存储器速度分配450来生成GPU存储器速度转换单元326。

[0141] GPU存储器速度分配450对应于用户可访问的GPU存储器的时钟速度。例如,仅物理GPU存储器的时钟速度的一部分可被分配给一个或更多个用户的工作负载。换句话说,被分配的GPU存储器的可用时钟速度可测量小于物理GPU存储器的默认时钟速度。GPU存储器速度计算部分452通常可以对应于表示由工作负载110中的一个或更多个使用的GPU存储器速度资源的权重。在一些情况下,GPU存储器计算部分452可以是实验推导的值,该实验推导的值是基于对若干用户的工作负载的分析的。照此,在一些情况下,GPU存储器速度计算部分452通常可以被描述为动态权重,该动态权重在预定时间是可调整的以应对工作负载的变化。

[0142] GPU处理器计算部分444、GPU存储器计算部分448和GPU存储器速度计算部分452的和可以总计达100。换句话说,通过将计算部分除以100,每个计算部分可表示为由第二资源

转换单元生成器404评估的总计算资源的百分比。

[0143] 如先前所讨论的,物理计算资源108可附加地或替代地包括其他资源209。如所讨论的,其他资源209可包括现场可编程门阵列(FPGA)、张量处理单元(TPU)、智能处理单元(IPU)、神经处理单元(NPU)、视频处理单元(VPU)、数字信号处理器(DSP)、加速处理单元(APU)、专用集成电路(ASIC)、片上系统(SoC)、可编程SoC、专用标准部件(ASSP),自适应计算加速平台(ACAP)、微控制器和/或任何其他计算资源中的任何一个或更多个。在这些情况下,可使用与其他计算资源209中的相应一个对应的独立资源转换单元生成器来确定每个资源转换单元。例如,转换单元生成器400可以包括其他资源转换单元生成器454。其他资源转换单元生成器454可被配置为生成与其他资源209中的相应一个或更多个对应的一个或更多个转换单元。

[0144] 图5示出了用于将不同计算资源的使用量化为单个测量单元的方法500的示例流程图,该方法可以例如在图1的系统中使用。方法500可实施为一个或更多个非暂时性计算机可读介质(例如,储存器206)中的一个或更多个指令,所述一个或更多个指令被配置为由一个或更多个处理器(例如,CPU202)执行。

[0145] 如图所示,方法500可以包括步骤502。步骤502可包括使工作负载(例如,工作负载110)在计算系统(例如,计算系统104)上执行。在一些情况下,例如,响应于从设备102接收到请求,可以使得工作负载被执行。另外地或可替代地,工作负载可以包括由计算系统的操作引起的一个或更多个空闲进程。

[0146] 方法500还可以包括步骤504。步骤504可以包括测量在一段时间(例如,1秒、10秒、1分钟、1小时、1天、1周、1个月、1年和/或任何其他合适的时间段)内由工作负载的执行所使用(例如,消耗)的物理资源的量。可测量(与例如虚拟资源相对的)资源的物理量以防止或以其他方式缓解由膨胀技术使用的不准确的读数,该膨胀技术可以由测量虚拟资源使用的管理程序采用。进一步,通过直接测量物理资源,可以在多种形式的硬件上实现该方法,包括不运行虚拟机器的硬件的。因此,可以在利用封装软件(例如,容器)的系统中实现该方法。

[0147] 方法500可以包括步骤506。步骤506可以包括对所使用的物理资源的测量值进行归一化。将测量值归一化可包括将不同资源(例如,CPU资源、存储器资源、储存器资源、网络接口资源、GPU资源和/或任何其它计算资源)中的每一个转换成通用单元(例如,如关于图3和图4所描述的)。

[0148] 方法500可以包括步骤508。步骤508可以包括对归一化的测量值求和。方法500还可以包括步骤510。步骤510可以包括生成表示所使用的被求和的归一化资源的单个值。例如,所生成的单个值可以是求和结果。在其他情况下,所生成的值可以表示在时间窗上被求和的归一化资源的平均值(例如,每周平均值,其中各个被求和的归一化值对应于一周中的各天)。

[0149] 图6示出了用于将不同计算资源的最大或最小值量化为单个测量单元的方法600的示例流程图,该方法例如可以在图1的系统中使用。该方法可以被实现为一个或更多个非暂时性计算机可读介质(例如,储存器206)中的一个或更多个指令,该一个或更多个指令被配置为由一个或更多个处理器(例如,CPU202)执行。

[0150] 该方法可以包括步骤602。步骤602可以包括在一段时间内测量可用的物理资源的

最大量和/或在一段时间内测量用于运行空闲进程(例如,空闲进程113)的物理资源的最小量。

[0151] 该方法可以包括步骤604。步骤604可以包括归一化可用的不同物理资源

[0152] (例如,已使用和非活动资源)的测量值。将测量值归一化可包括将不同资源

[0153] (例如,CPU资源、存储器资源、储存器资源、网络接口资源、GPU资源和/或任何其它计算资源)中的每一个转换成通用单元(例如,如关于图3A-C和图4所描述的)。

[0154] 该方法还可以包括步骤606。步骤606可以包括对可用的物理资源的归一化测量值进行求和。该方法可以包括步骤608。步骤608可以包括生成表示被求和的可用的归一化资源的单个值。例如,所生成的单个值可以是求和结果。作为另一个示例,所生成的单个值可以是基于求和结果的调整后的值。

[0155] 在一些情况下,可以至少部分地基于与各个物理计算资源相对应的规范来计算在一段时间内可用的物理计算资源的最大和/或最小量。这可以允许用户估计可用的物理计算资源的最大和/或最小量,而无需访问物理计算资源。在一些情况下,在方法600中,可以将计算值与测量值进行比较。例如,可以将计算值与测量值进行比较以调整与各个计算资源和/或计算方法相对应的规范,使得可以改进对可用的物理计算资源的最大和/或最小量的估计。

[0156] 根据本公开的一个方面,提供了一种计算网络。计算网络可以包括被配置为执行至少一个工作负载的物理计算资源。计算网络还可以包括被配置为测量在一段时间内使用的物理计算资源的量的计量器。计量器可以生成表示在该段时间内使用的物理资源的量的单个使用值。

[0157] 根据本公开的另一方面,提供了至少一种计算机可读储存介质,其上存储有指令,当由一个或更多个处理器执行时,该指令使一个或更多个处理器执行一个或更多个操作。一个或更多个操作可以包括测量在一段时间内使用的一个或更多个物理计算资源的量。所述一个或更多个操作还可以包括归一化正在使用的每个物理计算资源的每个测量的量。所述一个或更多个操作还可以包括对正在使用的物理计算资源的归一化的测量的量进行求和。所述一个或更多个操作还可以包括基于被求和的归一化物理计算资源来生成表示在该段时间内使用的物理计算资源的单个使用值。

[0158] 根据本公开的又一方面,提供了一种方法。该方法可以包括测量在一段时间内正在被使用的一个或更多个物理计算资源的量。该方法还可以包括对正在被使用的每个物理计算资源的每个测量的量进行归一化。该方法还可以包括对正被使用的物理计算资源的归一化的测量的量求和。该方法还可以包括基于被求和的归一化物理计算资源来生成表示在该时间段内正在被使用的物理计算资源的单个使用值。

[0159] 根据本公开的又一方面,提供了一种计算网络。该计算网络可以包括被配置为通信地耦接到设备的计算系统。计算系统可以被配置为执行至少一个工作负载。计算系统可以包括被配置为执行至少一个工作负载的物理计算资源。计算系统还可以包括计量器,该计量器被配置为测量在一段时间内使用的物理计算资源的量。计量器可以生成表示在该时间段内使用的物理资源的量的单个使用值。

[0160] 根据本公开的又一方面,提供了一种用于计量不同计算机资源的系统。计量器可以包括至少一个收集器。收集器可以包括至少一个度量接收器,该度量接收器被配置为在

一段时间内从至少一个计算系统接收一个或更多个度量,其中,度量对应于物理资源使用。计量器还可以包括至少一个分析平台,该分析平台被配置为归一化所接收的度量中的每一个,并且对归一化的度量求和以生成表示在该时间段内的物理资源使用的单个使用值。

[0161] 根据本公开的又一方面,提供了一种计算网络。计算网络可以包括计算系统和计量器,所述计算系统具有被配置为执行一个或更多个工作负载的多个物理计算资源,所述计量器通信地耦接到计算系统。计量器可以被配置为测量在一段时间内使用的物理计算资源的量。计量器可以包括至少一个收集器,所述收集器包括至少一个度量接收器,所述度量接收器被配置为从计算系统接收一个或更多个度量,其中,度量对应于所使用的物理资源的量。计量器还可以包括至少一个分析平台,该至少一个分析平台与收集器通信并且被配置为归一化所接收的度量中的每一个并且对归一化的度量求和以生成表示在该时间段内的物理资源使用的单个使用值。

[0162] 根据本公开的另一方面,提供了一种计算网络。计算网络可包括被配置为通信地耦接到设备的计算系统。计算系统可以被配置成执行至少一个工作负载。该计算系统可以包括被配置成执行该至少一个工作负载的物理计算资源。计算系统可以被配置成与计量器通信。计量器可以被配置为测量正在使用的物理计算资源的量,其中,计量器还被配置为生成表示所使用的物理资源的量的单个使用值。

[0163] 根据本公开的另一方面,提供了一种计算网络。计算网络可包括具有被配置为执行一个或更多个工作负载的多个物理计算资源的计算系统。计算网络还可包括具有至少一个度量接收器的至少一个度量收集器,所述度量接收器配置成从计算系统接收一个或更多个度量,其中所述一个或更多个度量对应于物理计算资源的使用。计算网络还可包括具有备份接收器的至少一个备份收集器,备份接收器被配置为从计算系统接收一个或更多个度量。计算网络还可包括与度量收集器或备份收集器中的至少一个通信的至少一个分析平台,其中分析平台可被配置为归一化所接收的度量中的每一个并对经归一化的度量求和以生成单个使用值。

[0164] 根据本公开的另一方面,提供了一种用于计量不同计算机资源的系统。该系统可以包括至少一个收集器,该至少一个收集器包括至少一个度量接收器和账户管理器。度量接收器可配置成从至少一个计算系统接收一个或更多个度量,其中所述度量对应于物理资源使用。账户管理器可以被配置为管理多个账户。多个账户中的每个账户可以与至少一个资源消耗用户相关联,其中,每个资源消耗用户的物理资源使用与相应的账户相关联。该系统还可包括使用分析器,该使用分析器被配置成为一个或更多个账户生成表示该物理资源使用的每单元成本的基准。

[0165] 根据本公开的另一方面,提供了一种计算网络。计算网络可包括被配置为通信地耦接到设备的计算系统。计算系统还可以被配置为执行至少一个工作负载。该计算系统可以包括被配置成执行至少一个工作负载的物理计算资源。计算网络还可包括计量器,该计量器被配置成测量在一段时间内使用的物理计算资源的量并生成表示在该时间段内使用的物理资源的量的单个使用值。计量器可以进一步被配置为选择性地禁用设备对计算系统的至少一部分的访问。

[0166] 根据本公开的又一个方面,提供了一种方法。所述方法可以包括测量在一段时间内正在被使用的多个物理计算资源的量。所述方法还可包含归一化正在被使用的每个物理

计算资源的每个测量的量。该方法还可包括对正在被使用的物理计算资源的归一化的测量的量求和以生成表示在该时间段内正被使用的物理计算资源的单个使用值。

[0167] 在一些情况下,多个物理计算资源可包括多个不同物理计算资源。在一些情况下,所述多个物理计算资源可以包括图形处理系统、现场可编程门阵列、专用集成电路、片上系统、数字信号处理器、微控制器或自适应计算加速平台中的一个或更多个。在一些情况下,多个物理计算资源可以包括图形处理系统,以及图形处理系统可以包括图形处理器和图形存储器。在一些情况下,所述方法可进一步包含测量在所述时间段内可用的多个物理计算资源的最大量,将每个可用物理计算资源的每个测量的最大量归一化,以及对可用的物理计算资源的归一化的测量的最大量进行求和以生成表示在该时间段内可用的物理计算资源的最大量的单个可用资源值。

[0168] 根据本公开的另一方面,提供了一种计算网络。计算网络可包括具有配置成执行一个或更多个工作负载的多个物理计算资源的一个或更多个计算系统。计算网络还可包括一个或更多个度量收集器,每个度量收集器具有至少一个度量接收器,该度量接收器被配置为从一个或更多个计算系统接收多个度量。多个度量可对应于物理计算资源的使用。计算网络还可包括一个或更多个分析平台,所述分析平台被配置成归一化接收到的度量中的每一个。

[0169] 在一些情况下,一个或更多个分析平台可以进一步被配置成对归一化的度量进行求和以生成单个使用值。在一些情况下,物理计算资源使用的至少一部分可与一个或更多个帐户相关联。在一些情况下,一个或更多个帐户可包括多个帐户,至少一个帐户可与物理计算资源使用的一部分相关联,并且至少一个其他帐户可与物理计算资源使用的不同部分相关联。在一些情况下,度量收集器中的至少一个还可以包括具有一个或更多个插件的插件数据库,其中,一个或更多个插件可以被配置为使得度量收集器中的至少一个能够与一个或更多个计算系统通信。在一些情况下,分析平台可被配置为至少部分地基于物理计算资源使用与阈值的比较来生成通知。在一些情况下,该一个或更多个分析平台可包括多个分析平台,并且这些分析平台中的至少一个可以是备份分析平台。在一些情况下,一个或更多个分析平台可被配置为在预定时间激活或停用度量收集器中的至少一个。在一些情况下,度量收集器中的至少一个可包括一个或更多个分析平台中的至少一个。在一些情况下,一个或更多个度量收集器可以包括多个度量收集器,多个度量收集器中的至少一个可以是备份度量收集器。在一些情况下,分析平台可进一步被配置成至少部分地基于所接收的度量来确定对应于一段时间内的每个物理计算资源的高峰使用(例如,最大使用)、低谷使用(例如,最小使用)或平均使用中的一个或更多个。

[0170] 根据本公开的另一方面,提供了其上储存有一个或更多个指令的至少一个计算机可读储存介质,当被一个或更多个处理器执行时,该一个或更多个指令使得一个或更多个操作被执行。操作可以包括测量在一段时间内使用的多个物理计算资源的量。操作还可包括归一化正在被使用的每个物理计算资源的每个测量的量。操作还可包括对正在被使用的物理计算资源的归一化的测量的量进行求和以生成表示在该时间段内正在被使用的物理计算资源的单个使用值。

[0171] 在一些情况下,多个物理计算资源可包括多个不同物理计算资源。在一些情况下,所述多个物理计算资源可以包括图形处理系统、现场可编程门阵列、专用集成电路、片上系

统、数字信号处理器、微控制器或自适应计算加速平台中的一个或更多个。在一些情况下，多个物理计算资源可以包括图形处理系统，且图形处理系统可以包括图形处理器和图形存储器。在一些情况下，操作可进一步包含测量在一段时间内可用的多个物理计算资源的最大量，归一化每个可用的物理计算资源的每个测量的最大量，以及对可用的物理计算资源的归一化的测量的最大量进行求和以生成表示在一段时间段内可用的物理计算资源的最大量的单个可用资源值。在一些情况下，操作还可以包括访问插件数据库，其中，插件数据库可以包括被配置为使得能够测量多个物理计算资源的一个或更多个插件。在一些情况下，操作可还可包含确定对应于该段时间内每一物理计算资源的高峰使用(例如，最大使用)、低谷使用(例如，最小使用)或平均使用中的一个或更多个。

[0172] 在本公开的又一个方面中，提供了一种方法。该方法可以包括测量在一段时间内正在被使用的图形处理系统资源的量，其中，图形处理系统资源可以包括图形处理器和图形存储器。该方法还可以包括归一化与图形处理器和图形存储器相对应的被测量的量。该方法还可以包括对归一化的测量的量进行求和以生成表示在该时间段内正在被使用的图形处理系统资源的单个使用值。

[0173] 虽然已经在此描述和展示了本公开的若干实施例，但本领域的普通技术人员将容易地想到用于执行本文描述的功能和/或获得结果和/或一个或更多个优点的各种其他装置和/或结构，并且此类变化和/或修改均被视为在本公开的范围内。更一般地，本领域技术人员将容易理解，本文中所描述的所有参数、尺寸、材料和配置旨在是示例性的，并且实际的参数、尺寸、材料和/或配置将取决于使用本公开的教导的一个或更多个具体应用。

[0174] 本领域技术人员使用不超过常规实验将认识到或能够确定在本文描述的本公开的具体实施例的许多等效物。因此，应当理解，前述实施例仅以示例的方式呈现，并且在所附权利要求及其等同物的范围内，本公开可以以不同于具体描述和所要求的方式来实践。本公开针对在本文描述的每个单独的特征、系统、物品、材料、套件和/或方法。此外，如果此类特征、系统、物品、材料、套件和/或方法不相互矛盾，两个或更多个此类特征、系统、物品、材料、套件和/或方法的任何组合包括在本公开的范围内。

[0175] 除非明确指出相反，否则本文在说明书和权利要求书中使用的不定冠词“一”和“一个”应理解为是指“至少一个。”

[0176] 本文所使用的术语“耦接”和“被耦接”涵盖直接和间接耦接两者，除非明确指出相反。

[0177] 本文在说明书和权利要求书中所使用的词组“和/或”应当被理解为意指如此结合的元件中的“任一个或两者”，即，在一些情况下结合地存在并且在其他情况下分离地存在的元件。除非明确指示相反，除了由“和/或”子句具体标识的元件之外，可以可选地存在其他元件，无论是与具体标识的那些元件相关还是不相关。

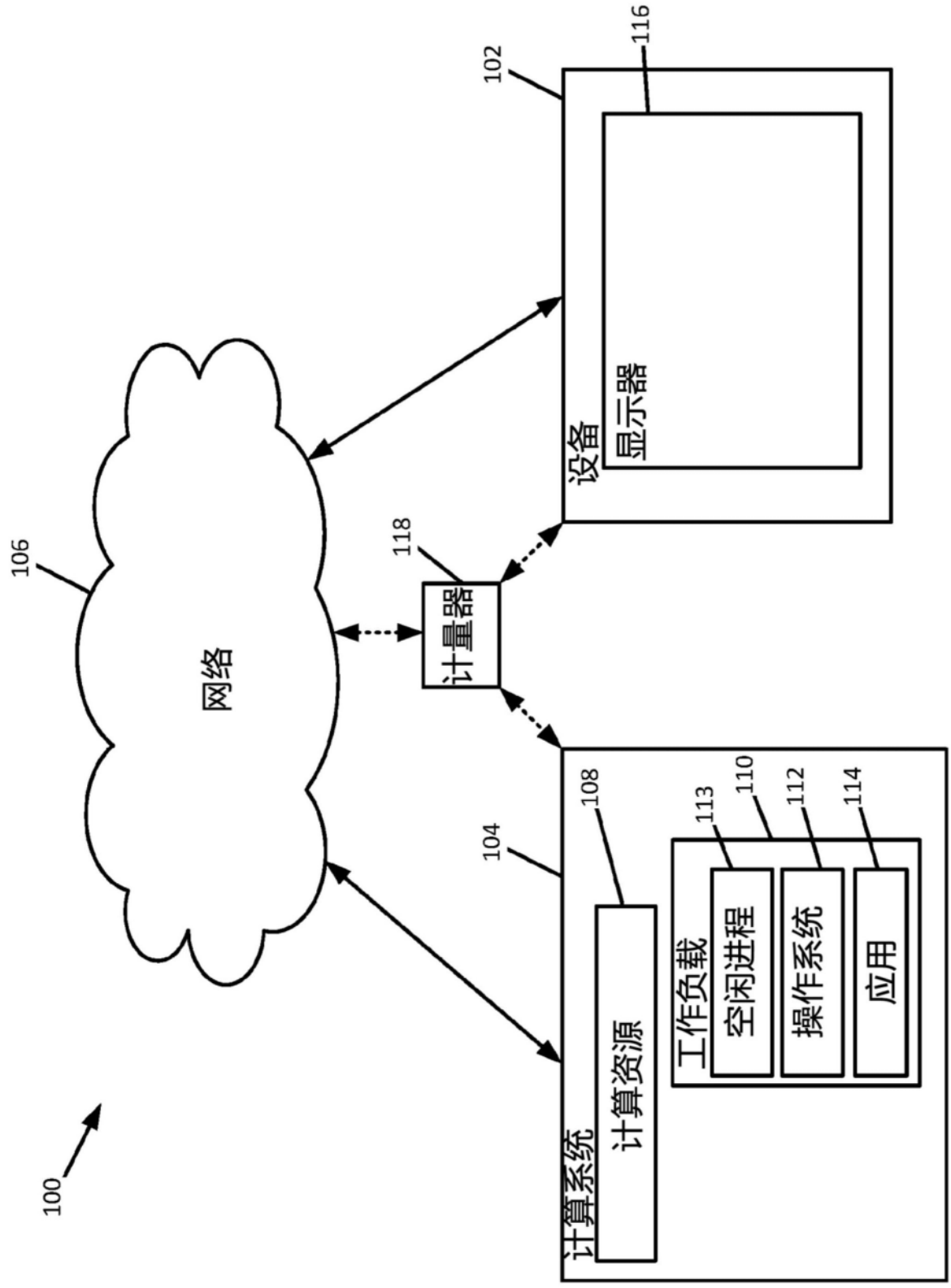

100
106
网络
102
116
设备
显示器
118
计量器
104
108
计算系统
计算资源
113
110
112
114
工作负载
空闲进程
操作系统
应用

图1

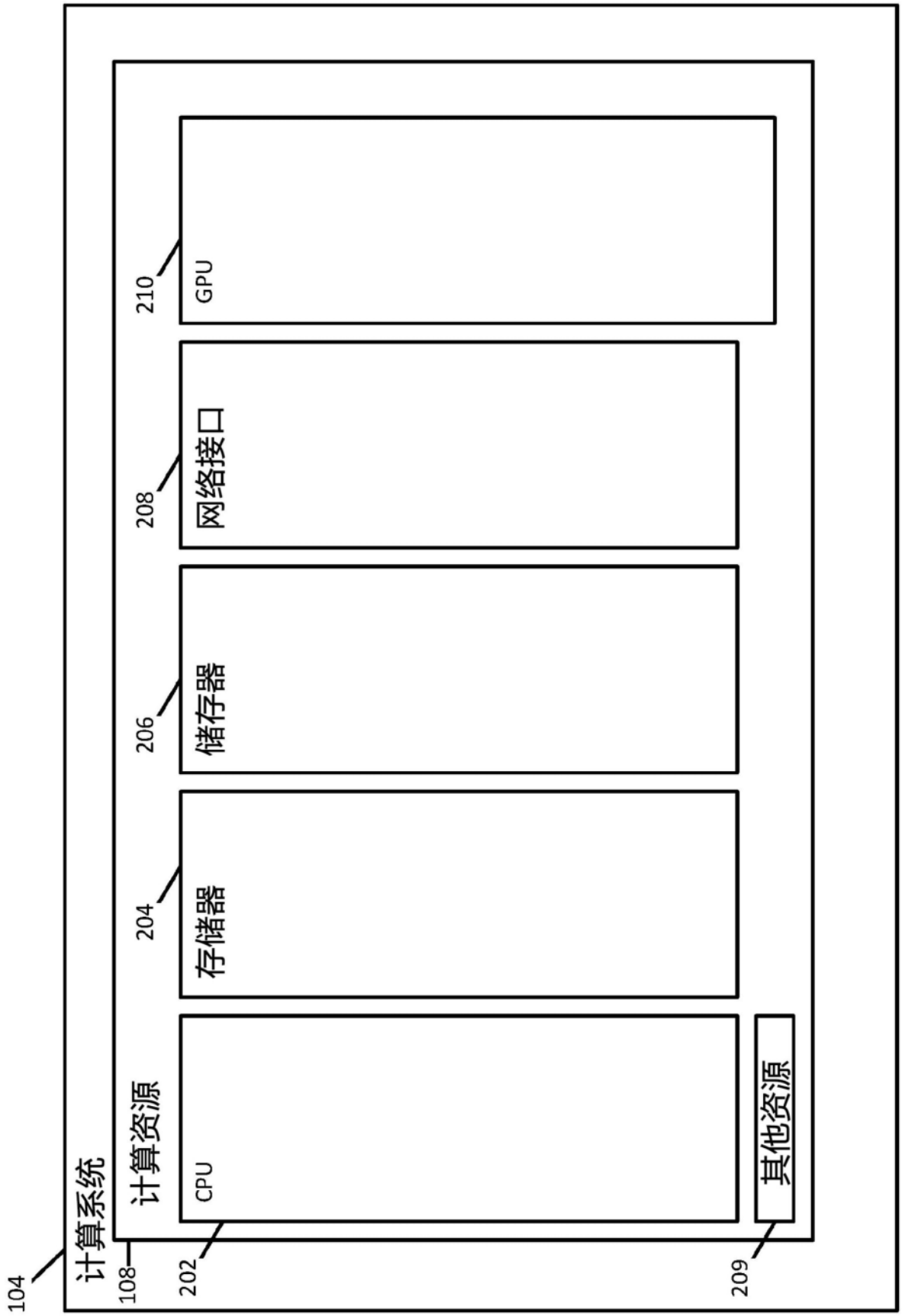

104
计算系统
108
计算资源
202
CPU
204
存储器
206
储存器
208
网络接口
210
GPU
209
其他资源

图2A

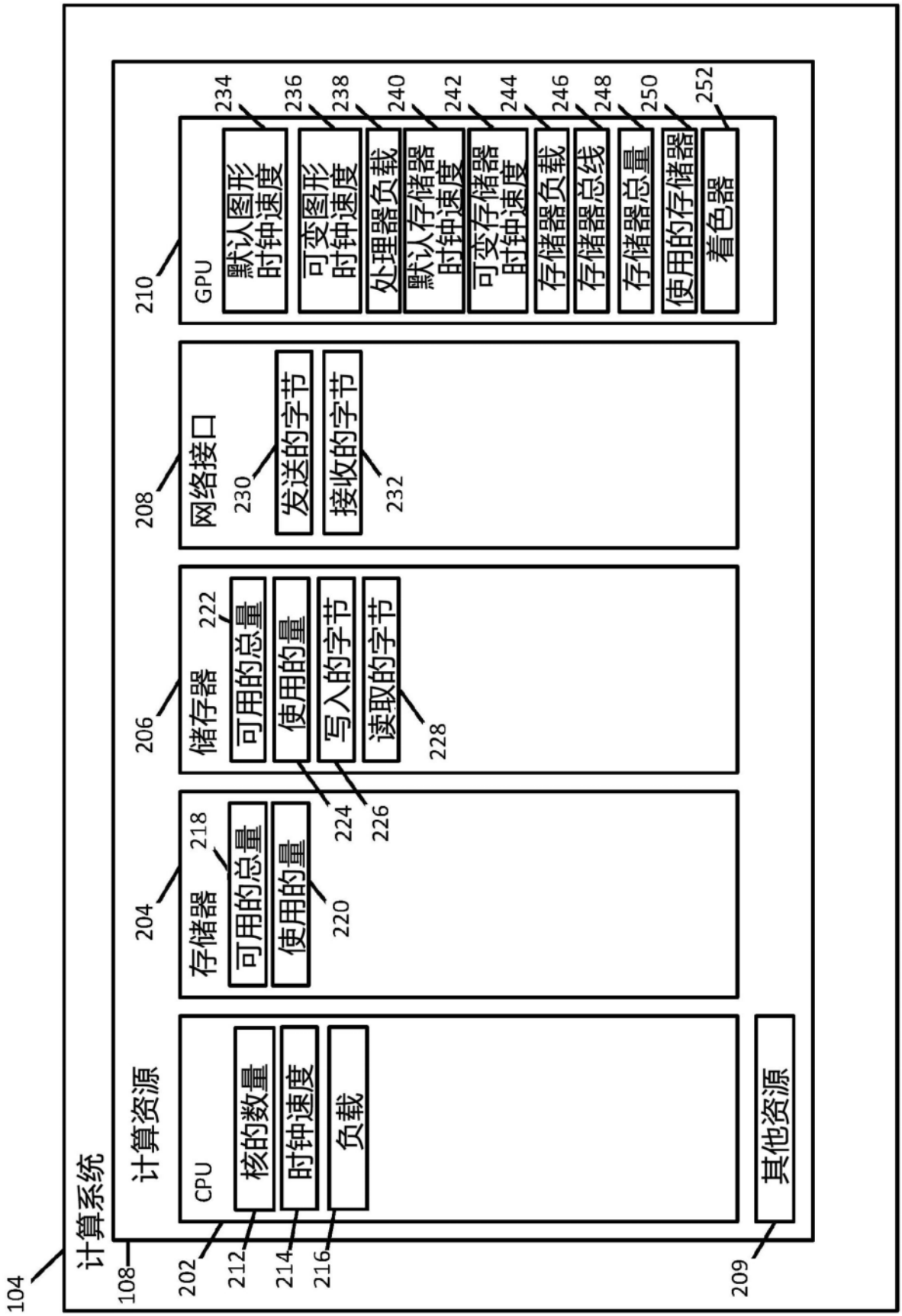

104
计算系统
108
计算资源
202
CPU
212
核的数量
214
时钟速度
216
负载
204
存储器
218
可用的总量
220
使用的量
206
储存器
222
可用的总量
224
使用的量
226
写入的字节
228
读取的字节
208
网络接口
230
发送的字节
232
接收的字节
210
GPU
234
默认图形时钟速度
236
可变图形时钟速度
238
处理器负载
240
默认存储器时钟速度
242
可变存储器时钟速度
244
存储器负载
246
存储器总线
248
存储器总量
250
使用的存储器
252
着色器
209
其他资源

图2B

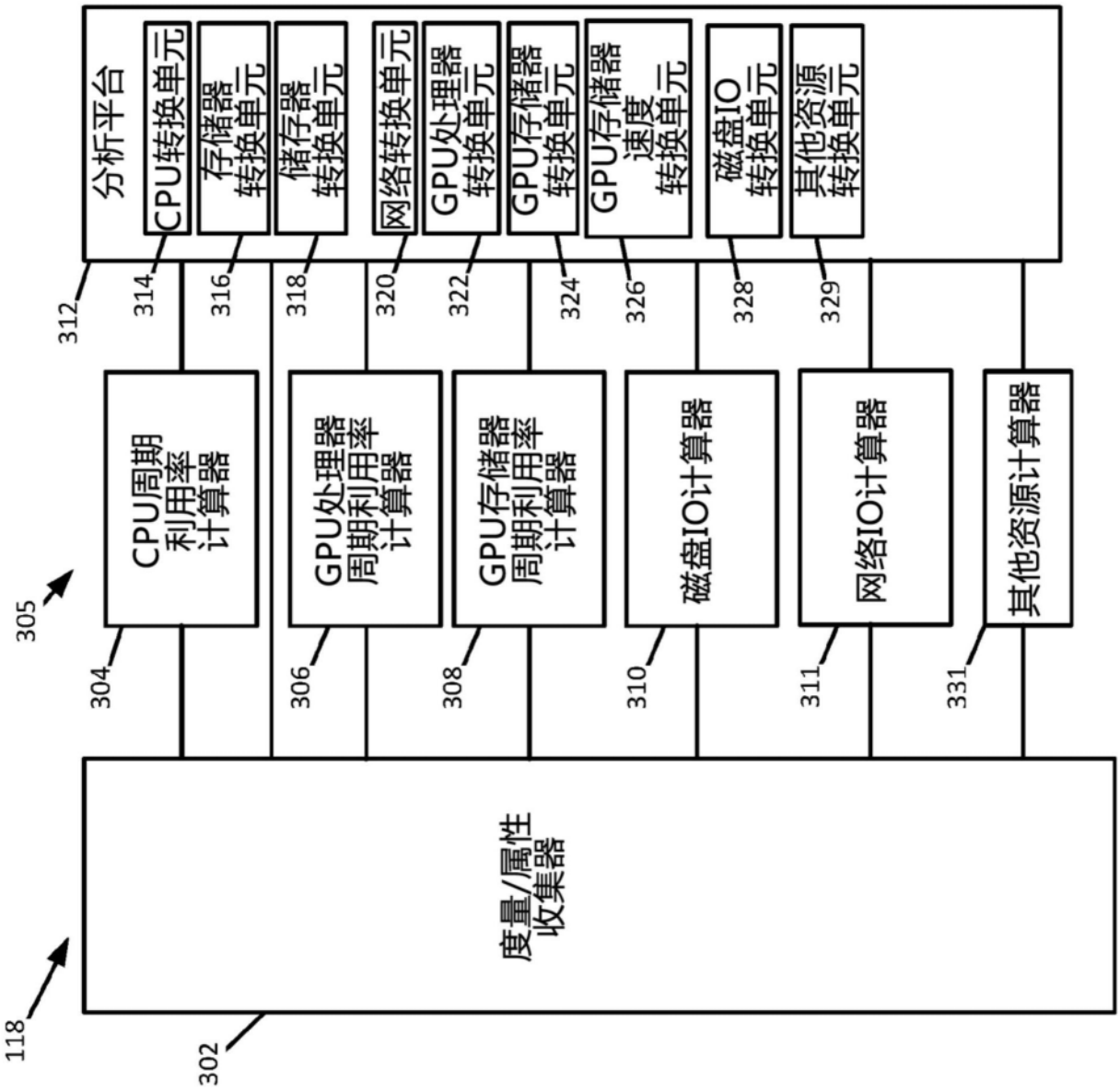

分析平台
CPU转换单元
存储器转换单元
储存器转换单元
网络转换单元
GPU处理器转换单元
GPU存储器转换单元
GPU存储器速度转换单元
磁盘IO转换单元
其他资源转换单元
312
314
316
318
320
322
324
326
328
329
CPU周期利用率计算器
GPU处理器周期利用率计算器
GPU存储器周期利用率计算器
磁盘IO计算器
网络IO计算器
其他资源计算器
304
306
308
310
311
331
305
度量/属性收集器
118
302

图3A

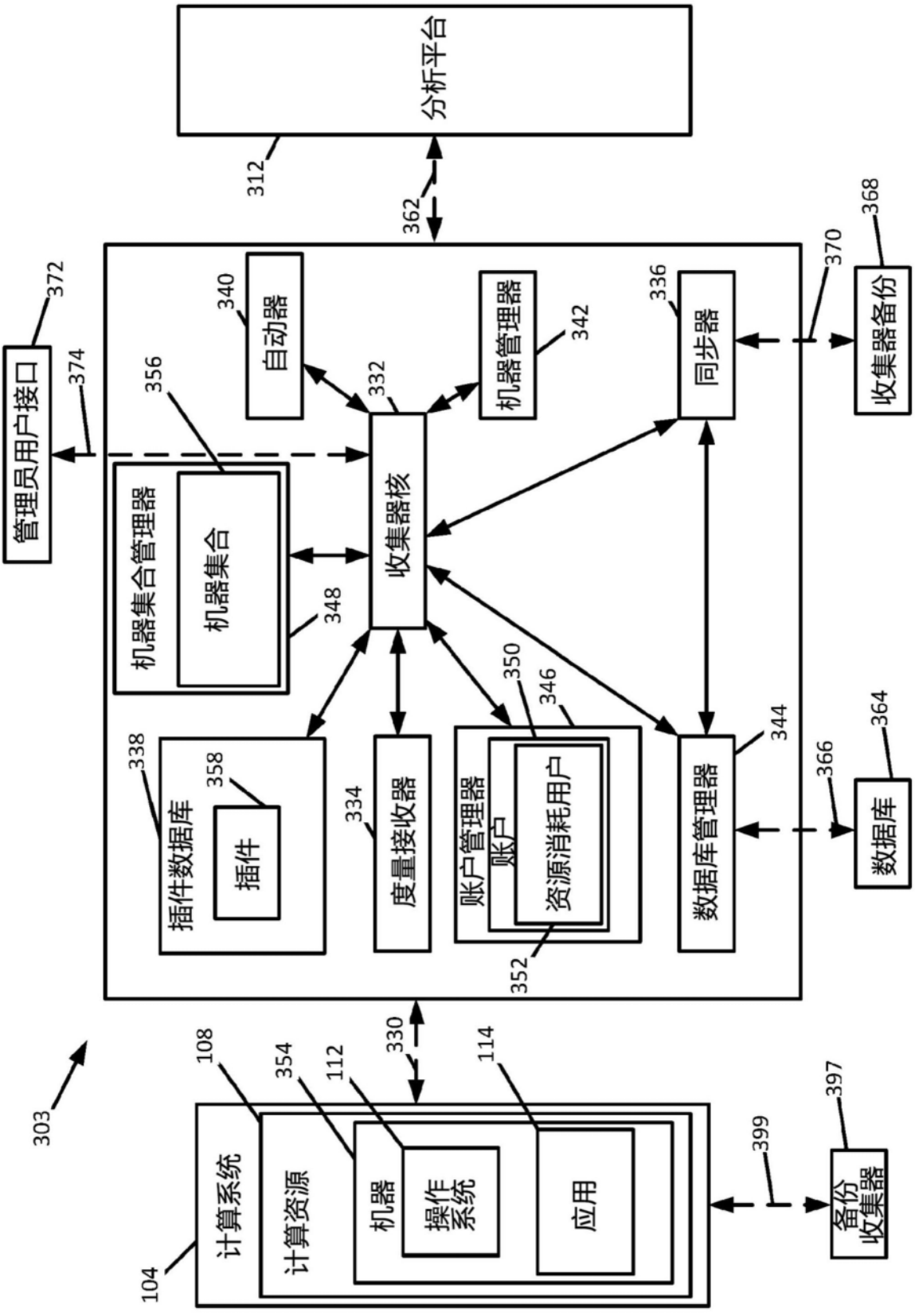

分析平台
312
362
368
收集器备份
370
372
管理员用户接口
374
340
自动器
332
机器管理器
342
336
同步器
356
机器集合管理器
机器集合
348
收集器核
插件数据库
338
358
插件
度量接收器
334
账户管理器
账户
资源消耗用户
350
346
352
数据库管理器
344
366
364
数据库
303
108
354
112
330
114
计算系统
计算资源
机器
操作系统
应用
104
399
397
备份收集器

图3B

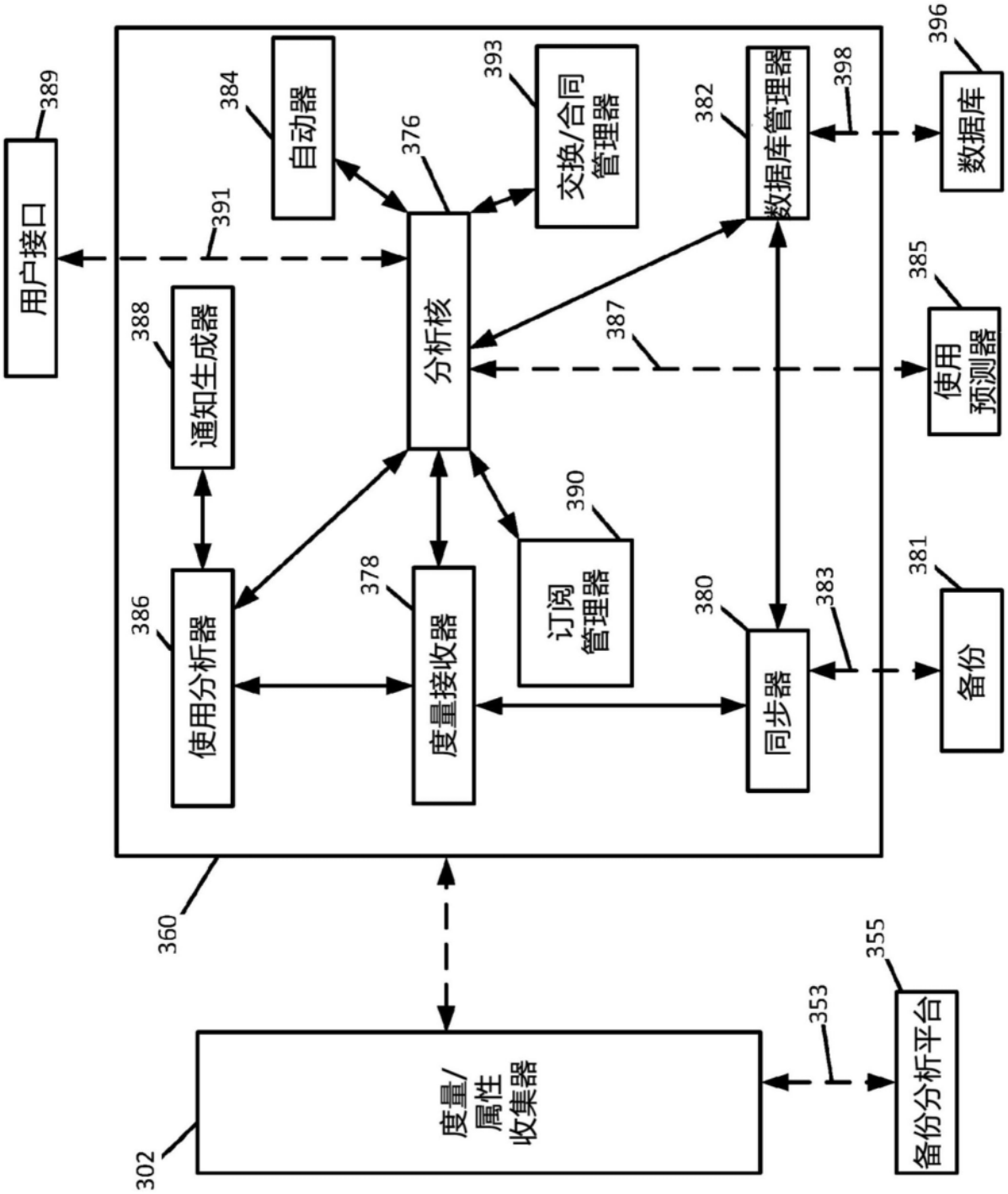

用户接口
389
自动器
384
交换/合同管理器
393
数据库管理器
382
数据库
396
398
391
376
分析核
通知生成器
388
387
使用预测器
385
订阅管理器
390
使用分析器
386
度量接收器
378
同步器
380
383
备份
381
360
355
353
度量/属性收集器
302
备份分析平台

图3C

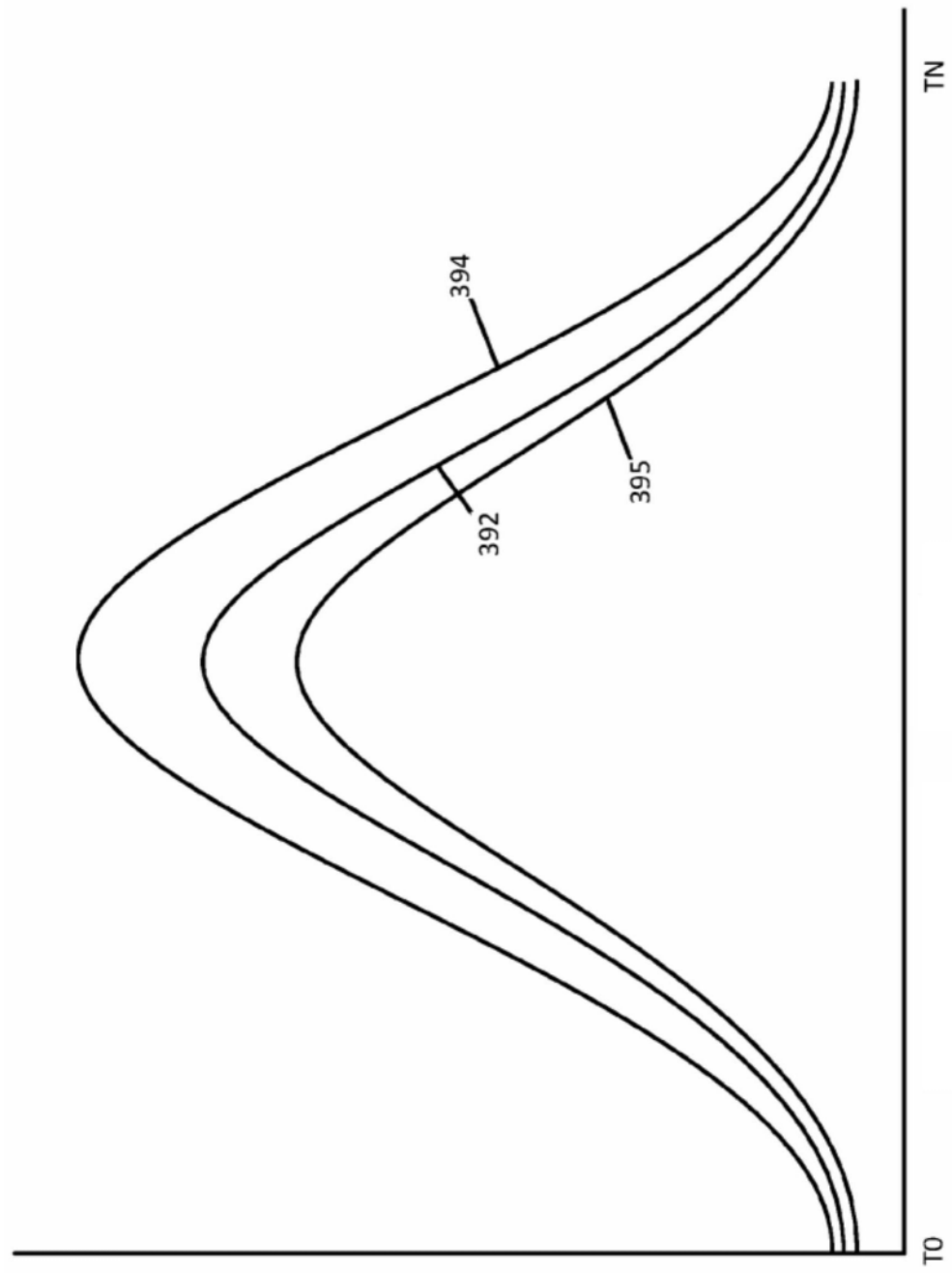


图3D

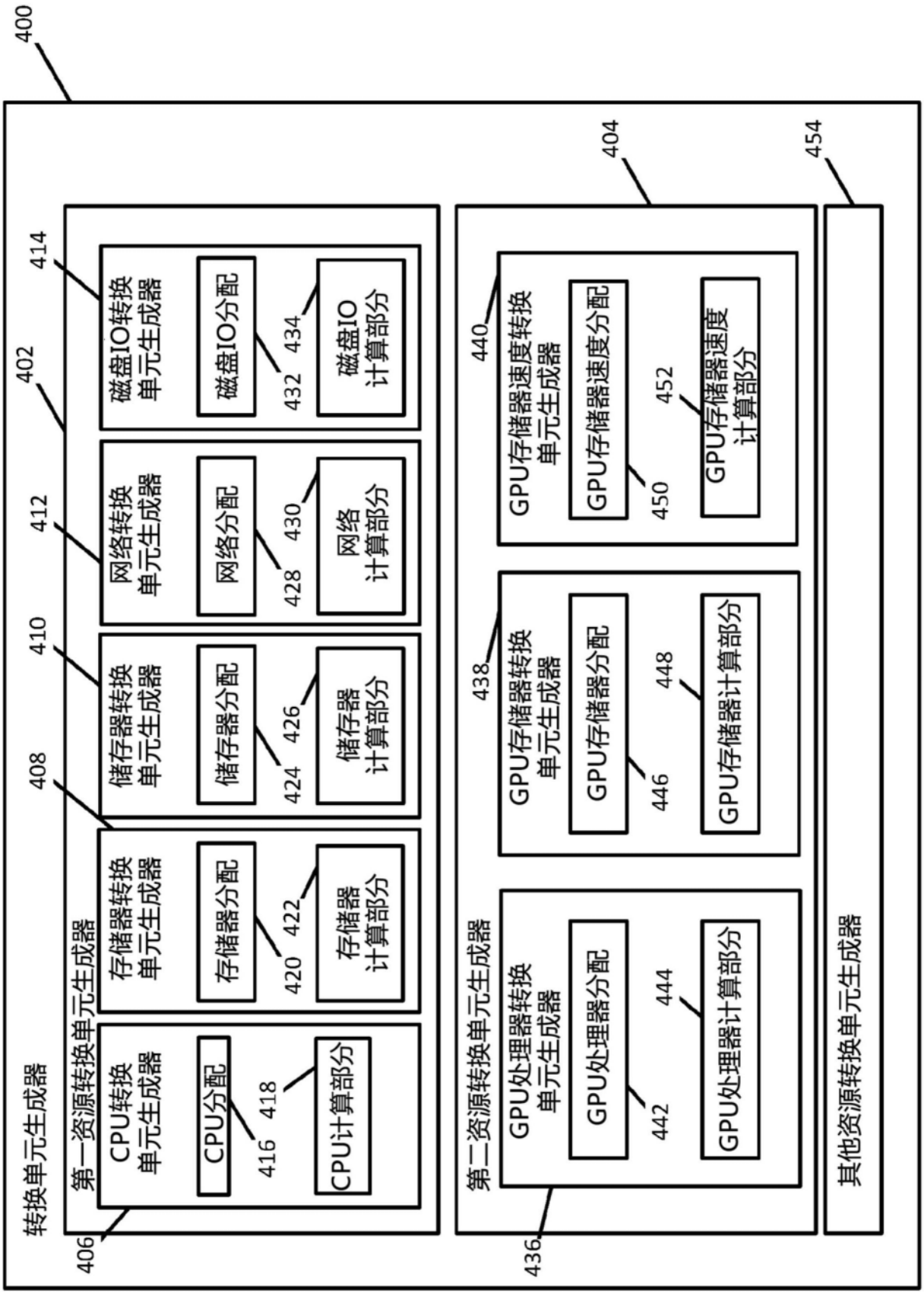

400
转换单元生成器
第一资源转换单元生成器
406
408
CPU转换
单元生成器
CPU分配
416
418
CPU计算部分
410
存储器转换
单元生成器
存储器分配
420
422
存储器
计算部分
412
网络转换
单元生成器
网络分配
428
430
网络
计算部分
402
414
磁盘IO转换
单元生成器
磁盘IO分配
432
434
磁盘IO
计算部分
第二资源转换单元生成器
436
438
GPU处理器转换
单元生成器
GPU处理器分配
442
444
GPU处理器计算部分
GPU存储器转换
单元生成器
GPU存储器分配
446
448
GPU存储器计算部分
404
440
GPU存储器速度转换
单元生成器
GPU存储器速度分配
450
452
GPU存储器速度
计算部分
454
其他资源转换单元生成器

图4

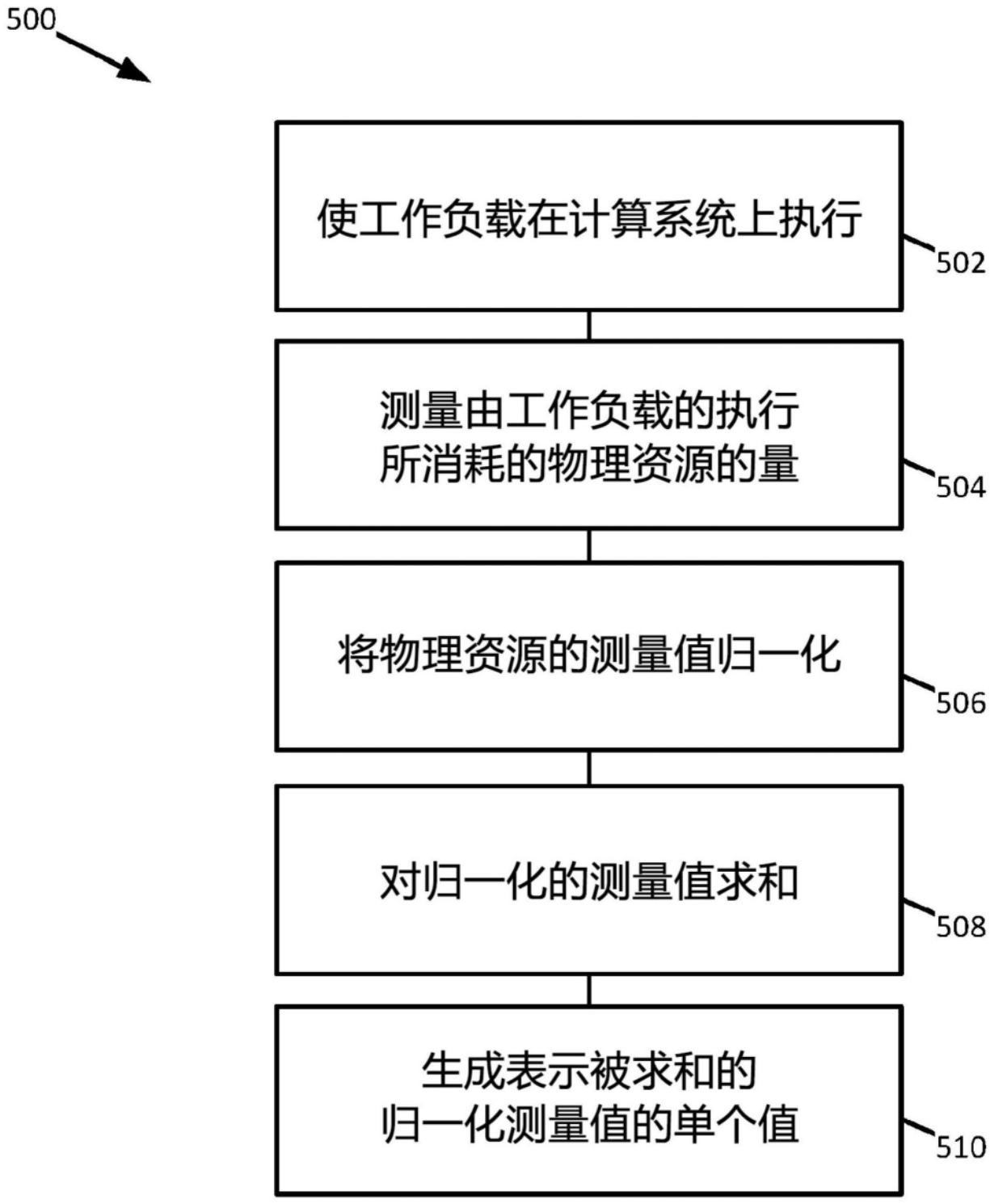

500
使工作负载在计算系统上执行
502
测量由工作负载的执行
所消耗的物理资源的量
504
将物理资源的测量值归一化
506
对归一化的测量值求和
508
生成表示被求和的
归一化测量值的单个值
510

图5

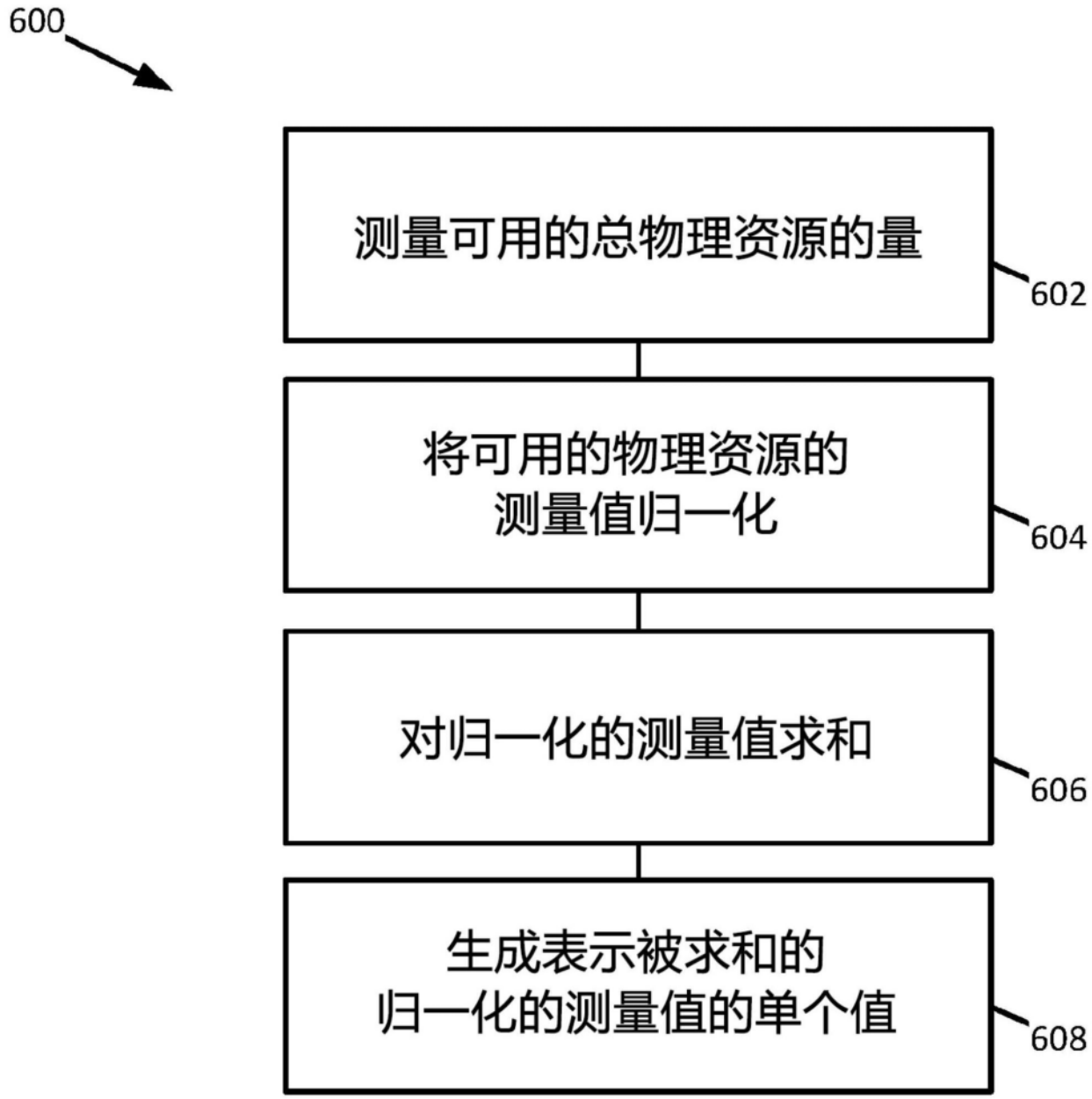

600
测量可用的总物理资源的量
602
将可用的物理资源的
测量值归一化
604
对归一化的测量值求和
606
生成表示被求和的
归一化的测量值的单个值
608

图6