



(12) 发明专利申请

(10) 申请公布号 CN 104572294 A

(43) 申请公布日 2015. 04. 29

(21) 申请号 201410558606. 7

(22) 申请日 2014. 10. 20

(30) 优先权数据

14/057, 898 2013. 10. 18 US

(71) 申请人 奈飞公司

地址 美国加利福尼亚州

(72) 发明人 丹尼尔·艾萨克·雅克博逊

尼尔拉杰·乔希 谱尼特·欧勃莱

咏·袁 菲利普·西蒙·图菲丝

(74) 专利代理机构 北京东方亿思知识产权代理

有限责任公司 11258

代理人 李晓冬

(51) Int. Cl.

G06F 9/50(2006. 01)

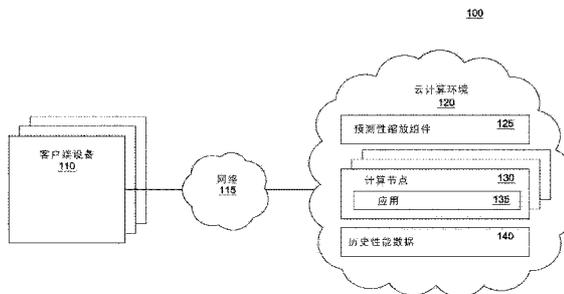
权利要求书3页 说明书9页 附图5页

(54) 发明名称

预测性自动缩放引擎

(57) 摘要

本公开提供了一种预测性自动缩放引擎,并描述了用于预测性的缩放分布式应用的技术。实施例可以在第一时间窗内监控云计算环境内的应用的性能以收集历史性能数据。在此,应用包括多个应用实例。可以在第二时间窗内监控应用的工作负载以收集历史工作负载数据。实施例可以分析历史性能数据和历史工作负载数据二者以确定应用的一个或多个缩放模式。在确定应用的当前状态匹配一个或多个缩放模式中的一个时,可以确定用于预测性地缩放应用的计划。随后,实施例可以基于所确定的计划来预测性地缩放多个应用实例。



1. 一种方法,包括:

在第一时间窗内监控云计算环境内的应用的性能以收集历史性能数据,其中所述应用包括多个应用实例;

在第二时间窗内监控所述应用的工作负载以收集历史工作负载数据;

分析所述历史性能数据和所述历史工作负载数据二者以确定所述应用的一个或多个缩放模式;

在确定所述应用的当前状态与所述一个或多个缩放模式中的一个缩放模式相匹配时,确定用于预测性地缩放所述应用的计划;

基于确定的计划来预测性地缩放所述多个应用实例;以及

在第二时间窗内监控具有经缩放的所述多个应用实例的所述应用的性能以收集附加性能数据和附加工作负载数据中的至少一者,其中所述附加性能数据和附加工作负载数据中的至少一者被用于影响未来缩放事件。

2. 如权利要求 1 所述的方法,其中确定用于预测性地缩放所述应用的计划还包括:

基于所述应用的当前工作负载和相匹配的缩放模式,确定所述应用的工作负载将在第一时间点处增加;

基于所述应用的当前工作负载确定所述应用的估计的未来工作负载;以及

确定用来实例化以满足所估计的未来工作负载的附加应用实例的数量。

3. 如权利要求 2 所述的方法,其中所述历史性能数据包括所述多个应用实例的启动时间数据,

其中确定用于预测性地缩放所述应用的计划还包括:

基于所述启动时间数据来确定用于实例化该数量的附加应用实例的第一时间线,

其中预测性地缩放所述多个应用实例还基于所确定的第一时间线,并且还包括:

根据所确定的第一时间线来实例化一个或多个附加应用实例。

4. 如权利要求 3 所述的方法,其中确定用于预测性地缩放所述应用的计划还包括:

基于所述应用的当前状态和所述相匹配的缩放模式,确定所述应用的工作负载将在第一时间点处减少;

基于所述应用的当前工作负载确定所述应用的估计的未来工作负载,其中所估计的未来工作负载少于所述应用的当前工作负载;以及

确定处理所估计的未来工作负载将需要的所述多个应用实例的数量,其中所确定的应用实例的数量少于应用实例的当前数量。

5. 如权利要求 4 所述的方法,其中所述历史性能数据包括所述多个应用实例的关闭时间数据,

其中确定用于预测性地缩放所述应用的计划还包括:

基于所述关闭时间数据来确定用于关闭应用实例从而达到所确定的应用实例的数量的第二时间线,

其中预测性地缩放所述多个应用实例还基于所确定的第二时间线,并且还包括:

根据所确定的第二时间线来关闭所述多个应用实例中的一个或多个。

6. 如权利要求 1 所述的方法,其中所述历史性能数据至少包括对于每单位时间由应用实例处理的请求的度量,并且其中所述历史工作负载数据至少包括对于每第二单位时间的

到来请求的度量。

7. 如权利要求 1 所述的方法,其中分析所述历史性能数据和所述历史工作负载数据二者以确定所述应用的一个或多个缩放模式还包括:

确定所述应用的当前状态的一个或多个属性,该一个或多个属性指示了在随后的时间点处所述应用的未来工作负载。

8. 如权利要求 7 所述的方法,其中所述一个或多个属性包括以下各项中的至少一项:星期几、当日时间、所述应用的当前工作负载、所述应用的当前工作负载的增加速率、所述应用的当前工作负载的减少速率、以及促销事件日程表。

9. 如权利要求 1 所述的方法,还包括:

分析所述历史性能数据以检测一个或多个异常,其中所述一个或多个异常代表通常并不指示所述应用的日常性能的独立事件,

其中分析所述历史性能数据和所述历史工作负载数据二者以确定所述应用的一个或多个缩放模式还包括:为了所确定的所述一个或多个缩放模式的目的,忽视所述历史性能数据内的检测到的一个或多个异常。

10. 如权利要求 1 所述的方法,其中所述应用的当前状态包括以下各项中的至少一项:当前星期几、当前时间值、所述应用的当前工作负载、所述应用的当前工作负载的当前改变速率、所述多个应用实例中的应用实例的当前数量、以及所述应用的事件日程表中的条目。

11. 如权利要求 1 所述的方法,其中确定用于预测性地缩放所述应用的计划还包括:

当确定所述应用的当前状态偏离所述历史性能数据多于阈值偏离量时:

确定所述应用的当前状态代表性性能异常;

基于所述历史性能数据来确定所述应用的估计的未来工作负载;以及

基于所估计的未来工作负载而不基于所述应用的当前工作负载来确定用于预测性地缩放所述应用的计划。

12. 如权利要求 1 所述的方法,其中预测性地缩放所述多个应用实例是结合另一种应用缩放技术和负载均衡算法中的至少一者来进行操作的。

13. 如权利要求 12 所述的方法,其中基于所确定的计划来预测性地缩放所述多个应用实例还包括:

基于所确定的计划来确定要维护的应用实例的阈值数量;以及

在不将所述多个应用实例缩放至低于所述应用实例的阈值数量的情况下,响应于所述应用的当前工作负载来反应性地缩放所述多个应用实例。

14. 一种系统,包括:

处理器;以及

含有程序的存储器,所述程序当在所述处理器上执行时执行包括以下各项的操作:

在第一时间窗内监控云计算环境内的应用的性能以收集历史性能数据,其中所述应用包括多个应用实例;

在第二时间窗内监控所述应用的工作负载以收集历史工作负载数据;

分析所述历史性能数据和所述历史工作负载数据二者以确定所述应用的一个或多个缩放模式;

在确定所述应用的当前状态与所述一个或多个缩放模式中的一个缩放模式相匹配时,

确定用于预测性地缩放所述应用的计划；

基于确定的计划来预测性地缩放所述多个应用实例；以及

在第二时间窗内监控具有经缩放的所述多个应用实例的应用的性能以收集附加性能数据和附加工作负载数据中的至少一者，其中所述附加性能数据和附加工作负载数据中的至少一者被用于影响未来缩放事件。

预测性自动缩放引擎

技术领域

[0001] 实施例一般涉及工作负载管理,并且更具体地涉及在预期到应用的未来工作负载的情况下预测性地缩放云计算环境内执行的应用的多个实例。

背景技术

[0002] 高效资源分配是对现代云计算环境的持续挑战。例如,特定应用可能需要 100 个应用实例来处理其峰值工作负载,但是可能仅需要 40 个应用实例来用于处理其平均工作负载。在此示例中,应用可以被配置为使用 40 个应用实例来操作,但是此配置将会不能适应应用的峰值工作负载。同样,应用可以被配置为一直使用 100 个应用实例来操作,但是此配置将会导致资源的低效使用,因为应用实例可能在非峰值工作负载期间处于空闲或未被充分利用。因此,云解决方案可能试图反应性地缩放应用实例的数量来满足波动的工作负载。也就是说,逻辑可以确定应用资源何时处于空闲或者应用何时不能跟上其当前的工作负载,并且可以将应用实例的数量相应地缩小或放大。然而,由于应用的工作负载在实践中几乎是确定地动态的,缩放应用实例的数量以适应增加或减少的工作负载是富有挑战性的。

[0003] 作为另一挑战,许多应用的启动过程需要大量时间。因此,如果附加的应用实例直到系统检测到该应用相对于当前的工作负载表现不佳时才被创建,并且应用实例花费大量时间来进行初始化,那么应用可能在等待应用实例被初始化的同时远远落后于其当前的工作负载。此情景可能导致应用的进一步表现不佳,并且在一些情况下可能导致系统反应性地产生甚至更多的附加应用实例来解决表现不佳,从而导致过量的应用实例。此外,在这样的反应性缩放技术不能足够快速地满足正增加的工作负载的情况下,系统可能完全不能赶上,从而潜在地导致中断、错误甚至整个系统故障。

发明内容

[0004] 本文公开的一个实施例提供了一种方法,该方法包括:在第一时间窗内监控云计算环境内的应用的性能以收集历史性能数据。应用包括多个应用实例。方法还包括在第二时间窗内监控应用的工作负载以收集历史工作负载数据。此外,方法包括分析历史性能数据和历史工作负载数据二者以确定应用的一个或多个缩放模式。该方法还包括:在确定应用的当前状态与一个或多个缩放模式中的一个相匹配时,确定用于预测性地缩放应用的计划。此外,方法包括:基于所确定的计划来预测性地缩放多个应用实例以及在第二时间窗内监控具有经缩放的多个应用实例的应用的性能以收集附加性能数据和附加工作负载数据中的至少一者,其中附加性能数据和附加工作负载数据中的至少一者被用于影响将来的缩放事件。

[0005] 其他实施例包括但不限于:包括使得处理单元能够实现所公开的方法的一个或多个方面的指令的计算机可读介质以及被配置为实现所公开的方法的一个或多个方面的系统。

附图说明

[0006] 因此,以可以详细理解本公开的上述特征的方式,可以参照实施例进行以上简要概述的更具体描述,一些实施例在附图中示出。然而,应注意,附图仅示出典型实施例,且因此不应认为限制其范围,因为本公开可以包括其他均等有效的实施例。

[0007] 图 1 示出根据本文描述的一个实施例的、配置有预测性缩放组件的计算基础设施。

[0008] 图 2 示出根据本文描述的一个实施例的、配置有预测性缩放组件的计算基础设施。

[0009] 图 3 是示出了根据本文描述的一个实施例的、用于确定云应用的性能模式的方法的流程图。

[0010] 图 4 是示出了根据本文描述的一个实施例的、用于预测性地缩放云应用的实例的方法的流程图。

[0011] 图 5 示出根据本文描述的一个实施例的、配置有预测性缩放组件的计算基础设施。

具体实施方式

[0012] 一般来说,许多大规模应用包括多个不同的应用实例。应用的工作负载可以通过(一个或多个)负载均衡算法的使用来分布至这些应用实例中。通常,以此方式设计应用存在许多优点。例如,可以添加或移除应用实例以动态地缩放应用的大小,从而允许应用处理增加的工作负载(即,通过添加实例)或释放空闲或未充分利用的计算资源(即,通过移除实例)。此外,如果一个应用实例出现故障,则剩余的应用实例可以承担出现故障的实例的工作负载,从而为应用提供冗余。

[0013] 管理在云环境中运行的大规模应用(例如,视频流服务及其支持应用)的一个挑战在于,此应用的工作负载可能随时间显著地变化。例如,这样的视频流服务可能在晚上在峰值电视观看时段期间经历其峰值工作负载,而相同的流服务可能在观看者的需求不那么大的清晨时段经历较轻的工作负载。此外,随着视频流服务增加新的观看者,一周中的给定夜晚的峰值工作负载将根据观看者的增长以基本上每周为基础增加。这可能在观众快速增长的时间期间(例如,在用于视频流服务的促销周期期间)特别明显。

[0014] 当前,云计算环境可以被配置为基于在给定时间点处应用的当前工作负载来反应性地缩放应用实例的数量。例如,云计算环境可以检测到当前实例化的应用实例不能跟上到来的请求的数量,并且因此可以确定应实例化附加的应用实例来帮助满足应用的当前工作负载。

[0015] 虽然此反应性缩放技术在某些情况下(例如,当应用的启动时间相对短时,当工作负载以相对慢的速率改变时等等)可以足够地执行,但是这些技术并不适用于一些应用。例如,特定应用的应用实例可能花费一个小时来完全实例化。在这样的应用的情况下,如果直到应用实例的当前数量不能跟上当前工作负载时才创建附加应用实例(即,反应性缩放解决方案),那么到来的请求的积压可能继续积累一个小时,直到将附加的应用实例带上线。这样,用户可能在该启动时间期间体验到延迟或服务中断。此外,由于工作负载在此

冗长的启动时间期间可能继续增加,所以等到附加应用实例被完全初始化的时候,附加应用实例的数量可能不再够用。此问题是复合的,因为在附加应用实例的启动时间期间积聚的请求积压也必须被处理。

[0016] 因此,实施例提供用于预测性地缩放用于云计算环境中的应用的多个应用实例的技术。实施例可以在第一时间窗内监控云计算环境内的应用的性能以收集历史性能数据。例如,实施例可以收集诸如应用实例在每单元时间可以处理多少请求以及用于创建新的应用实例的平均启动时间之类的性能信息。此外,实施例可以在第二时间窗内监控应用的工作负载以收集历史工作负载数据。在此,实施例可用收集关于应用的工作负载在一天内如何变化以及工作负载在一星期中的各天如何变化的信息。

[0017] 实施例随后可以分析历史性能数据和历史工作负载数据二者以确定针对应用的一个或多个缩放模式 (pattern)。例如,实施例可以对收集到的数据执行统计分析以确定指示了应用的工作负载的增加或减少的一个或多个模式。此外,实施例可以确定应用的当前状态与一个或多个缩放模式中的一个相匹配,并且作为响应可以确定用于预测性地缩放应用的计划。例如,实施例可以确定在未来时间点处可能被需要的多个应用实例,并且可以建立用于缩放当前数量的应用实例以在未来的时间点处实现所确定数量的应用实例的时间线。实施例随后可以基于所确定的计划来预测性地缩放多个应用实例。有利地是,这样做提供了能够对应用的工作负载的波动进行预期并且可以相应地缩放应用实例的数量的预测性缩放解决方案。

[0018] 图 1 示出了根据本文描述的一个实施例的、配置有预测性缩放组件的计算基础设施。如图所示,系统 100 包括多个客户端设备 110 和云计算环境,它们通过网络 115 连接。云计算通常是指作为网络上的服务的可缩放计算资源的配设。更正式来说,云计算可以被定义为这样的计算能力:在计算资源与其底层技术架构(例如,服务器、存储设备、网络)之间提供抽象化从而实现对可配置的计算资源的共享池的方便、按需的网络访问,该可配置的计算资源能够用最少的管理工作或服务提供商交互而被快速配设和释放。因此,云计算允许用户访问“云”中的虚拟计算资源(例如,存储设备、数据、应用、甚至完整的虚拟计算系统),而不考虑用来提供计算资源的底层物理系统(或者那些系统的位置)。通常,云计算资源是按使用计费 (pay-per-use) 地提供给用户,其中用户仅为实际使用的计算资源(例如,用户消耗的存储空间的数量或用户实例化的虚拟系统的数量)付费。用户可以在任何时间并且从互联网中的任何地方访问居于云中的资源。

[0019] 在此,云计算环境 120 包括预测性缩放组件 125、多个计算节点 130 以及历史性能数据 140。通常,计算节点 130 反映云计算环境 120 内的计算系统(例如,虚拟机)。在此,每个计算节点 130 包括分布式应用的应用实例 135。为了此示例的目的,假设托管在计算节点 130 上的应用实例 135 全部对应于单个分布式应用(例如,视频流服务)。应注意,虽然在每个计算节点 130 内示出单个应用实例 135,但是此描述仅用于说明性目的,并且其他实施例可以在计算节点 130 上实例化两个或更多个应用实例 135。更一般来说,本文描述的实施例可以配置具有用于将应用实例 135 托管在云计算环境 120 内的任何配置来进行使用。

[0020] 在此,预测性缩放组件 125 可以在某一时间段内监控应用实例 135 的执行以收集历史性能数据 140。通常,历史性能数据 140 包括与应用实例 135 的性能有关的数据,例如在单位时间期间特定应用实例平均处理多少到来的请求(例如,每分钟处理的请求)以及

那些请求的等待时间（例如，处理单个请求平均花费多久）。此外，预测性缩放组件 125 可以在某一时间段内监控分布式应用的工作负载。例如，预测性缩放组件 125 可以监控来自客户端设备 110 的多个到来的请求以收集用于分布式应用的历史工作负载数据。预测性缩放组件 125 随后可以执行工作负载数据的统计分析（并且也潜在地执行历史性能数据 140 的分析）从而确定一个或多个缩放模式。

[0021] 通常，缩放模式代表指示应用的未来工作负载的到来的工作负载的模式或趋势。这样的统计分析也可以考虑诸如当日时间、星期几、有关应用的工作负载等等的因素来确定缩放模式。例如，预测性缩放组件 125 可以确定对于给定的一天在峰值工作负载的时间之前的数个小时期间到来的工作负载的增加速率指示了峰值工作负载的量。例如，预测性缩放组件 125 可以确定在历史峰值工作负载的时间之前的数个小时期间（例如，当对视频流服务的需求最大的夜晚小时数）该应用的工作负载的相对大的增加速率指示该特定的一天的峰值工作负载将相对较高。作为另一示例，如果预测性缩放组件 125 确定应用的工作负载的增加速率相对较低，则预测性缩放组件 125 可以确定峰值工作负载对于特定的该天来说也将相对较低。当然，缩放模式的这些示例在非限制地提供并且仅用于说明性目的，并且更一般来说，与本文描述的功能相一致的、根据统计分析收集到的工作负载数据所确定的任何模式能够被使用。

[0022] 在确定缩放模式后，预测性缩放组件 125 可以监控应用的当前工作负载以确定当前工作负载何时匹配所确定的缩放模式中的一个。继续以上示例，预测性缩放组件 125 可以确定对于给定的一天，应用的工作负载的增加速率在历史峰值工作负载的时段之前的数个小时期间相对较高。预测性缩放组件 125 随后可以基于缩放模式和应用的当前工作负载确定应用的预期的未来工作负载。在一个实施例中，预测性缩放组件 125 被配置为基于应用的启动时间来确定在未来时间点处的应用的未来工作负载。例如，如果应用的启动时间是约 1 小时，则预测性缩放组件 125 可以确定从当前时间开始 1 小时的应用的未来工作负载。在一个实施例中，未来时间点是基于应用的平均启动时间加上一些预定义的时段（例如，10 分钟）来确定，以帮助确保将在未来工作负载之前启动应用实例。

[0023] 一旦估计出应用的未来工作负载，则预测性缩放组件 125 随后可以确定满足估计的未来工作负载所需的应用实例的数量。通常，如果未来工作负载大于应用的当前工作负载，则可能需要实例化多个附加的应用来适应增加的工作负载。另一方面，如果估计出的未来工作负载小于当前工作负载，则可以关闭一些数量的应用实例以避免空闲或未充分利用的计算资源。

[0024] 预测性缩放组件 125 然后可以基于所确定的满足估计出的未来工作负载所需要的应用实例的数量和应用实例的平均启动时间来确定用于预测性地缩放应用的计划。例如，预测性缩放组件 125 可以确定将需要 10 个附加应用实例来适应估计出的从现在开始 1 小时后的未来工作负载，并且还可以确定将需要 20 个附加的应用实例来适应估计出的从现在开始 2 小时后的未来工作负载。如果预测性缩放组件 125 随后确定应用实例具有 45 分钟的平均启动时间，则预测性缩放组件 125 可以确定计划来从现在开始 15 分钟后实例化 10 个附加的应用实例（即，以满足到估计出的未来 1 小时的工作负载）并且从现在开始 75 分钟后实例化 10 个附加的应用实例（即，满足估计出的未来 2 小时的工作负载）。有利地，这样做允许预测性缩放组件 125 基于应用的估计出的未来工作负载来预测性地缩放分

布式云应用,从而允许应用更平滑地处理工作负载的波动。

[0025] 另一方面,如果预测性缩放组件 125 确定未来工作负载小于当前工作负载并且将需要减少 10 个应用实例来处理未来工作负载,则预测性缩放组件 125 然后确定在时间窗内关闭特定应用实例的计划。这样做,预测性缩放组件 125 可以考虑应用实例的平均关闭时间。通过响应于预测出的未来工作负载的减少来减少应用实例的数量,预测性缩放组件 125 有助于避免未充分利用的应用实例,由此减少运行应用的成本。此外,这样做还释放与关闭的应用实例相关的计算资源,由此将这些资源释放以在需要时用于其他目的。

[0026] 通常,预测性缩放组件 125 也可以与其他缩放和负载管理技术组合,从而确保应用的系统健康、可用性和最优化。作为示例,在一个实施例中,预测性缩放组件 125 被配置为结合反应性缩放组件来进行操作,该反应性缩放组件被配置为响应于工作负载(例如,到来的请求容量)的改变来添加或移除应用实例。例如,预测性缩放组件 125 可以被配置为基于应用的估计出的未来工作负载提供最小数量的活动的应用实例,而反应性缩放组件可以响应于未预期的工作负载来实例化附加的应用实例。在此,虽然传统的反应性缩放技术会响应于请求容量的暂时下降来减少活动的应用实例的数量,这样做会在应用的正常请求容量恢复时导致服务故障。在此情景下,预测性缩放组件 125 可以基于应用在未来时间点处预测出的工作负载来防止应用实例的数量降到应用实例的阈值数量之下,从而避免在正常请求容量恢复时的服务故障。有利地,将预测性缩放组件 125 与其他缩放技术(诸如反应性缩放组件)相组合可以提升系统的性能,允许反应性缩放组件响应工作负载的波动同时基于应用的预测的未来工作负载确保最小数量的应用实例保持实例化。

[0027] 应注意,虽然以上示例涉及与反应性缩放组件相结合地工作的预测性缩放组件 125,但是此示例是非限制地提供并且仅用于说明性目的。更一般说来,明白地预料到预测性缩放组件 125 可以结合多种其他缩放技术和更一般的工作负载管理技术来工作,从而提升系统的性能和效率。例如,预测性缩放组件 125 可以结合硬件缩放组件来工作,该硬件缩放组件被配置为基于应用的工作负载缩放执行应用实例的资源(例如,存储器、处理器等)。作为另一实例,预测性缩放组件 125 可以结合这样的组件来工作,该组件被配置为响应于应用的当前工作负载和/或预测的未来工作负载而缩放下游资源(例如,多个数据库实例、在其上执行数据库实例的硬件资源等)。

[0028] 图 2 示出了根据本文描述的一个实施例的、配置有预测性缩放组件的计算基础设施。如图所示,云计算环境 200 包括预测性缩放组件 125、未分配的计算资源 210 以及多个计算节点 215。在此,每个计算节点 215 配置有一个或多个应用实例 135 和一个或多个监控组件 225。在此实施例中,监控组件 225 可以监控其各自的应用实例 135 从而收集性能和工作负载数据。监控组件 225 随后可以将收集到的数据传输给预测性缩放组件 125,并且预测性缩放组件 125 可以使用这些数据来预测性地缩放应用实例 135 的数量和计算节点 215 的数量。

[0029] 例如,预测性缩放组件 125 可以确定应用实例 135 的当前工作负载与特定缩放模式相匹配,这指示应用的工作负载将在一些时间段内大体增加。响应于此确定,预测性缩放组件 125 可以使用未分配的计算资源 210 来分配多个附加的计算节点 215,并且可以在这些附加的计算节点上初始化多个新的应用实例 135。通常,预测性缩放组件 125 优选地确定在某个未来时间点处应用的未来工作负载(该未来时间点至少与应用实例的平均启动时间

一样长),从而确保应用实例可以是完全可操作的并且准备好在该未来时间点处处理未来工作负载。例如,如果应用实例花费平均 60 分钟来完全初始化,则预测性缩放组件 125 可以配置成预测到未来 75 分钟的未来工作负载并且创建足以处理此估计出的未来工作负载的多个附加应用实例,从而使得应用实例将在未来工作负载之前被初始化且可操作。

[0030] 在一个实施例中,预测性缩放组件 125 被配置为监控应用流量随时间的增长并且相应地调整其应用缩放。例如,视频流服务可能随时间持续增加其订户基数,并且使用视频流服务的订户数量可能逐月增加。因此,处理在一周中的给定时间点处到来的工作负载所需要的应用实例数量可能与订户基数的改变成比例地继续增加。例如,如果订户基数在一年期间内增加 20%,则在平均周二晚上五点到来的工作负载如今可能比一年前高约 20%。因此,预测性缩放组件 125 可以配置成在生成用于缩放应用实例的计划时考虑用户的增长(或减少)。当然,此示例是非限制地提供并且仅用于说明性目的。

[0031] 此外,虽然预测性缩放组件 125 可以基于用户的增长而变更缩放计划,但是此增长可能并不始终以成比例的方式相关。例如,虽然在一年期间内订阅的增长可能稍微增加,但是在具体一天和时间处的平均工作负载在同一年内可能由于其他因素而大大增加。例如,由于在过去的一年内对视频流服务添加更多内容,所以订户的观看模式也可能因此改变。因此,预测性缩放组件 125 可以将订户的增长考虑作为预测应用的未来工作负载和相应地确定用于应用实例的数量的计划的许多因素之一。

[0032] 此外,预测性缩放组件 125 可以在预测应用的未来工作负载时考虑已知事件。这些事件的实例包括夏令时、特定内容的可用性(例如,特定体育赛事的直播流)、假期等。这些事件可以编程的方式被规定至预测性缩放组件 125(例如,每年发生的事件,诸如假期、夏令时等),或者可以由用户手动地规定(例如,特定事件的直播流)。然而,更一般来说,预测性缩放组件 125 可以考虑与应用的工作负载具有任何相关性的任何事件。此外,取决于应用的类型,不同事件可能与应用具有不同的相关性。例如,如果应用是视频流服务,则假期可能指示由于用户在家中而非工作而使得应用的工作负载增加。另一方面,如果应用是用于商业的个人应用,则假期可能指示由于用户在特定的日期离开办公室而使得应用的工作负载减少。有利地,通过考虑这些事件,预测性缩放组件 125 可以更加精确且可靠地预测应用的未来工作负载并且相应地缩放应用实例。

[0033] 图 3 是示出了根据本文描述的一个实施例的、用于确定云应用的性能模式的方法的流程图。如图所示,方法 300 始于框 310,其中预测性缩放组件 125 监控(一个或多个)应用实例的性能和工作负载以收集历史性能数据。这些数据可以包括(但不限于)每单位时间的到来的请求的数量、每单位时间由给定应用实例处理的请求的数量、新应用实例的启动时间、应用实例的关闭时间等。此外,预测性缩放组件 125 监控云计算环境相对于应用的性能(框 315)。例如,除了应用实例的启动时间之外,可能存在配置和初始化在其上运行新应用实例的新的云计算节点的附加启动时间。

[0034] 预测性缩放组件 125 随后分析收集到的性能数据从而确定应用的缩放模式(框 320)并且方法 300 结束。例如,预测性缩放组件 125 可以确定历史工作负载数据指示应用的工作负载在每天下午 5 点之前显著增加,并且峰值工作负载的量与大约此时间处的增长的速率成比例。预测性缩放组件 125 随后可以使用此缩放模式以预测性地缩放多个应用实例,从而适应预测出的未来工作负载。

[0035] 在一个实施例中,预测性缩放组件 125 被配置为在反馈回路中操作,其中数据被持续地反馈到预测性缩放组件 125 以改进未来缩放操作。这些数据可以包括例如描述应用的工作负载的数据、描述应用的一个或多个缩放模式的数据、描述缩放操作对系统整体影响的数据等。有利地,通过将这些数据反馈回到预测性缩放组件 125 中,实施例可以更加精确地预测何时需要未来缩放操作并且也可以优化这些未来缩放操作。例如,这些数据可以用来确定其中配设了大量附加应用实例的特定缩放计划导致未料想到的副作用(例如,对一个或多个下游组件施加突然压力)。响应于此确定,预测性缩放组件 125 可以从历史缩放操作和有关数据中学习,并且预测性缩放组件 125 可以优化未来缩放操作(例如,通过限制实例化的附加应用实例的数量、通过将最优大小的缩放操作均匀地分布在一个时间段内等)以最小化未来的未料想到的副作用。

[0036] 图 4 是示出了根据本文描述的一个实施例的、用于预测性地缩放云应用的实例的方法的流程图。如图所示,方法 400 始于框 410,其中预测性缩放组件 125 为托管在云计算环境中的应用提供缩放模式数据。这些模式数据可以例如使用图 3 中示出并且以上论述的方法 300 来确定。预测性缩放组件 125 随后监控在云环境中执行的应用(框 414),从而检测何时满足缩放模式中的一个(框 420)。在此,如果应用的当前性能和工作负载不匹配缩放模式中的任一个,则方法 400 返回到框 414,其中预测性缩放组件 125 继续监控应用。

[0037] 相反,如果预测性缩放组件 125 检测到应用的当前性能和工作负载匹配缩放模式中的一个,则方法 400 进行到框 425,其中预测性缩放组件 125 基于缩放模式确定应用的预测到的未来工作负载。例如,继续以上给出的示例,其中历史工作负载数据指示应用的工作负载在每天下午 5 点之前显著地增加并且峰值工作负载的量与大约此时的增加速率成比例,预测性缩放组件 125 可以基于在下午 5 点之前应用的工作负载的当前增加速率来确定应用的估计出的未来工作负载。如以上所论述,预测性缩放组件 125 可以被配置为在至少与应用实例的平均启动时间一样长的某个未来时间点处的未来工作负载,以确保应用实例能够在未来工作负载之前被完全初始化。

[0038] 一旦预测性缩放组件 125 已经确定了所估计的未来工作负载,则预测性缩放组件 125 基于所估计的未来工作负载确定要创建的附加应用实例的数量(方框 430)。应注意,此示例假定所估计的未来工作负载大于应用的当前工作负载。然而,在其中未来工作负载小于应用的当前工作负载的其他示例中,预测性缩放组件 125 可以确定满足所估计的未来工作负载不再需要的多个应用实例。

[0039] 返回到本示例,预测性缩放组件 125 还确定用于创建附加应用实例的时间线(框 435)。在此,预测性缩放组件 125 可以配置成在创建时间线时考虑应用实例的平均启动,从而确保附加的应用实例在增加的未来工作负载的预期下将被完全初始化。预测性缩放组件 125 随后基于所确定的时间线来实例化附加的应用实例(框 440),并且方法 400 结束。

[0040] 图 5 示出了根据本文所述的一个实施例,配置有预测性缩放组件的计算基础设施。如图所示,环境 700 包括预测性缩放系统 705,该系统包括(而非限于)中央处理单元(CPU)702、网络接口 708、互连 710 以及系统存储器 712。CPU 702 取回并执行存储在系统存储器 712 中的编程指令。类似地,CPU 702 存储并取回居于系统存储器 712 中的应用数据。互连 710 辅助 CPU 702、输入/输出(I/O)设备接口 706、存储设备 704、网络接口 708 以及系统存储器 712 之间的传输(诸如编程指令和应用数据的传输)。I/O 设备接口 706

被配置为从用户 I/O 设备 722 接收输入数据。用户 I/O 设备 722 的示例可以包括一个或多个按钮、键盘、以及鼠标或其他定点设备。I/O 设备接口 706 还可以包括配置成生成电子音频输出信号的音频输出单元，并且用户 I/O 设备还可以包括配置成响应于电子音频输出信号而生成声音输出的扬声器。I/O 设备的另一个示例是通常代表用于生成图像以供显示的任何技术上可行的装置的显示设备。例如，显示设备可以是液晶显示 (LCD) 显示器、CRT 显示器或 DLP 显示器。显示设备可以是包括用于接收数字或模拟电视信号的广播或电缆调谐器的 TV。

[0041] 包括 CPU 702 以代表单个 CPU、多个 CPU、具有多个处理核心的单个 CPU 等。并且通常包括系统存储器 712 以代表随机存取存储器。存储设备 704 可以是磁盘驱动存储设备。尽管示出为单个单元，但是存储设备 704 可以是固定和 / 或可移除存储设备的组合，诸如固定磁盘驱动、软盘驱动、磁带驱动、可移除存储器卡或光存储设备、网络附加存储设备 (NAS) 或存储区域网 (SAN)。网络接口 708 被配置为通过通信网络传输数据，例如，传输来自客户端设备的上下文令牌和本地化的数字资产以及将由基于动态上下文的汇编程序所产生的数字内容的汇编变体返回到客户端设备。

[0042] 系统存储器 712 存储预测性缩放组件 125，该组件被配置为随时间监控应用的性能和工作负载，以用于在预期到未来工作负载的情况下预测性地缩放多个应用实例。如上所述，预测性缩放组件 125 可以在第一时间窗内监控云计算环境内的应用的性能以收集历史性能数据，其中应用包括多个应用实例。此外，预测性缩放组件 125 可以在第二时间窗内监控应用的工作负载以收集历史工作负载数据。预测性缩放组件 125 随后可以分析历史性能数据和历史工作负载数据二者以确定应用的一个或多个缩放模式。在确定应用的当前状态匹配一个或多个缩放模式中的一个时，预测性缩放组件 125 可以确定用于预测性地缩放应用的计划，并且可以使用该计划来预测性地缩放多个应用实例。

[0043] 以上，参照各种实施例。然而，应理解，本公开并不限于具体描述的实施例。相反，以上特征和元件的任何组合（无论是否与不同的实施例有关）可被料想到来实施和实践本文描述的功能。此外，尽管特定实施例可以实现优于其他可能的解决方案和 / 或优于现有技术的优点，但是特定优点是否由给定实施例实现并不限制本公开。因此，前述的方面、特征、实施例和优点仅是说明性的，并且除了在（一个或多个）权利要求中明确详述之外，并不是对所附权利要求的考虑元素或限制。同样，对“本发明”的任何参考不应解释为本文公开的任何发明主题的概括，并且除了在（一个或多个）权利要求中明确详述之外，不应认为是所述权利要求的元素或限制。

[0044] 虽然以上是针对各种实施例，但是在不脱离其基本范围的情况下可以设想出其他和另外的实施例。例如，可以在硬件或软件中或者在硬件与软件的组合中实施实施例。一个实施例可以被实施为用于与计算机系统一起使用的程序产品。程序产品的（一个或多个）程序定义了实施例的功能（包括本文描述的方法）并且可以包含在各种计算机可读存储介质上。说明性的计算机可读存储介质包括（但不限于）：(i) 在其上永久地存储信息的不可写存储介质（例如，计算机内的只读存储器设备，诸如可由 CD-ROM 驱动读取的 CD-ROM 磁盘、闪存存储器、ROM 芯片或任何类型的固态非易失性半导体存储器）；以及 (ii) 在其上存储了可变更信息的可写存储介质（例如，软盘驱动器或硬盘驱动器内的软盘或任何类型的固态随机存取半导体存储器）。这样的计算机可读存储介质在携带指令了本文描述的功能。

能的计算机可读指令时是本公开的实施例。

[0045] 因此,本公开的范围由所附权利要求来确定。

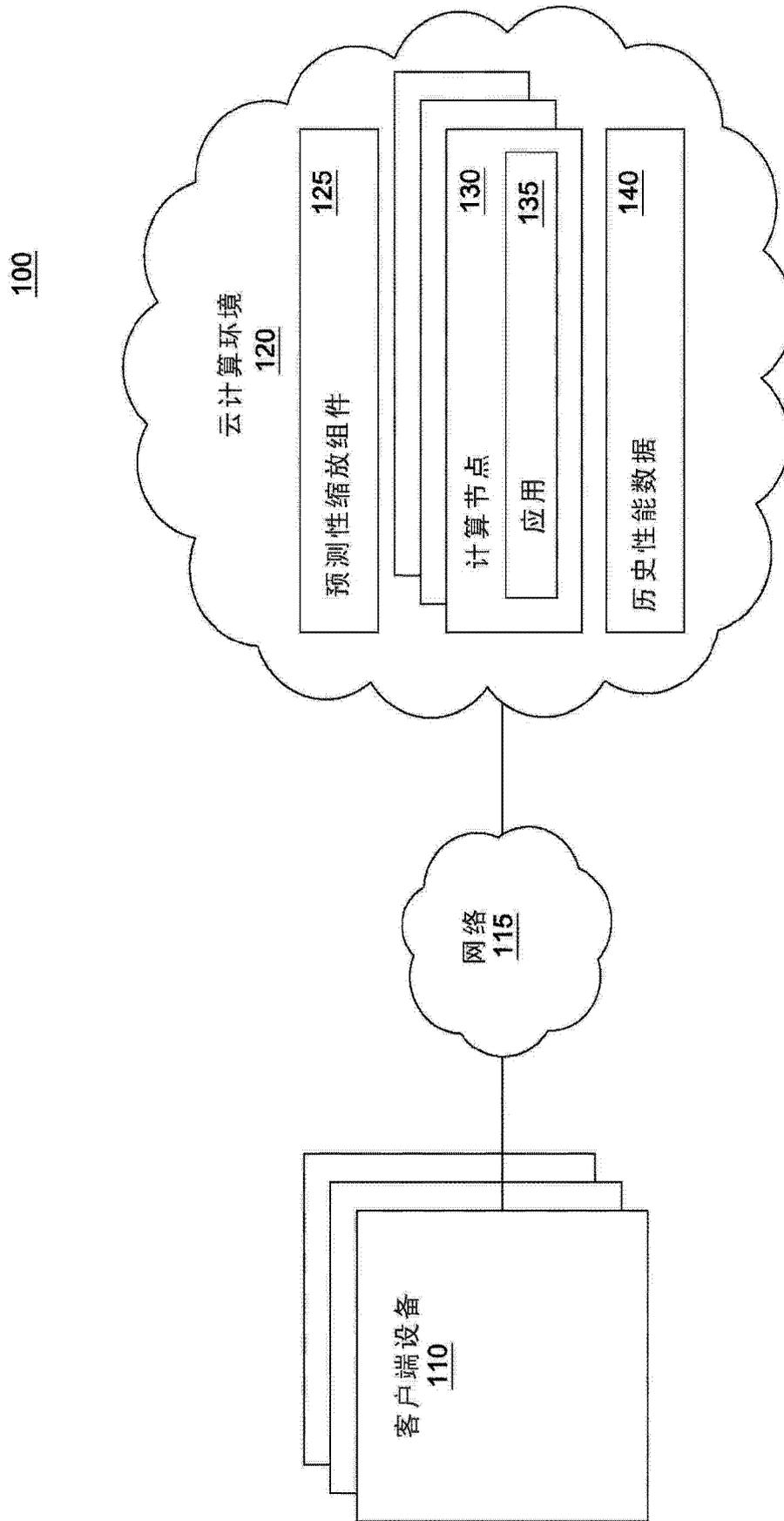


图 1

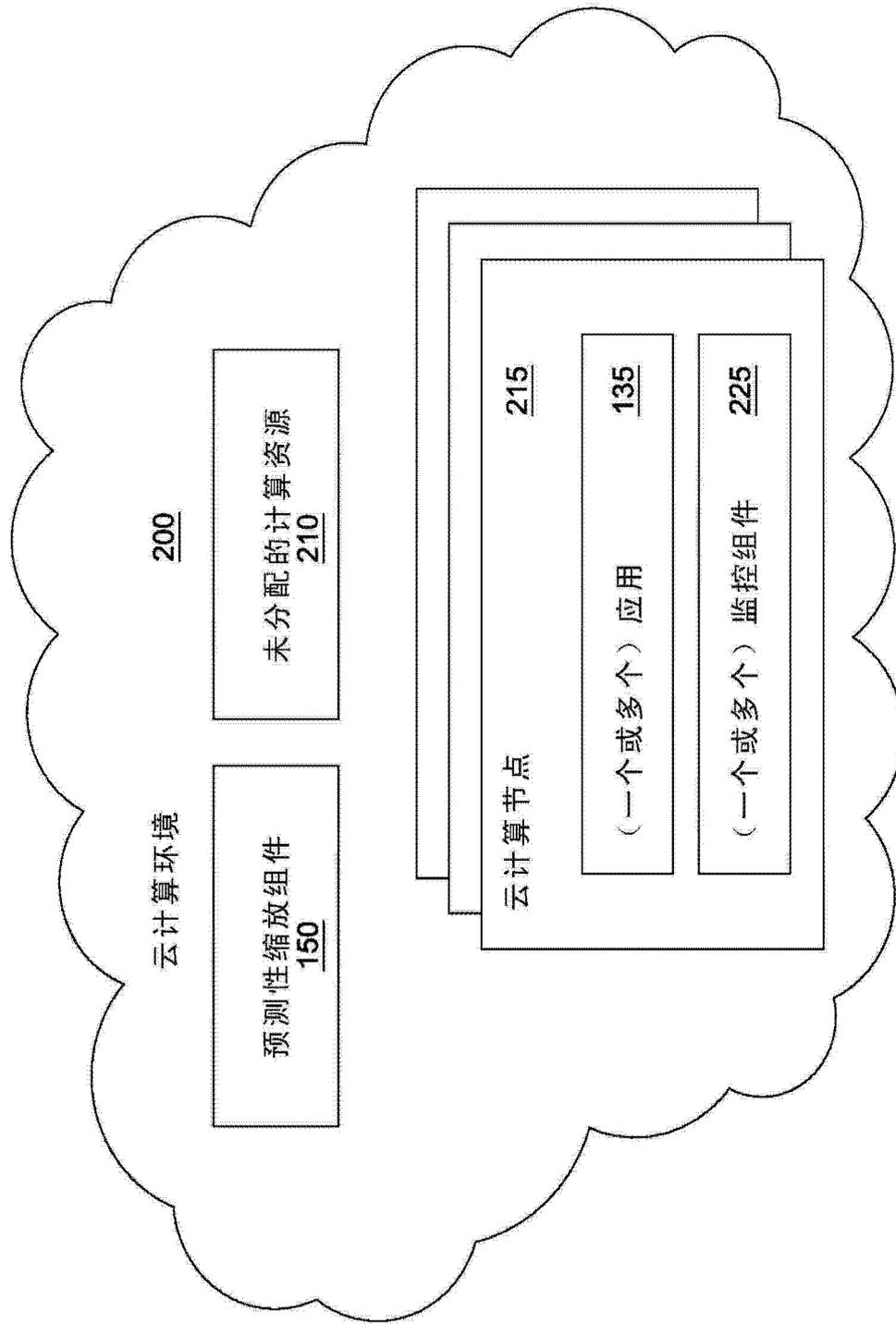


图 2

300

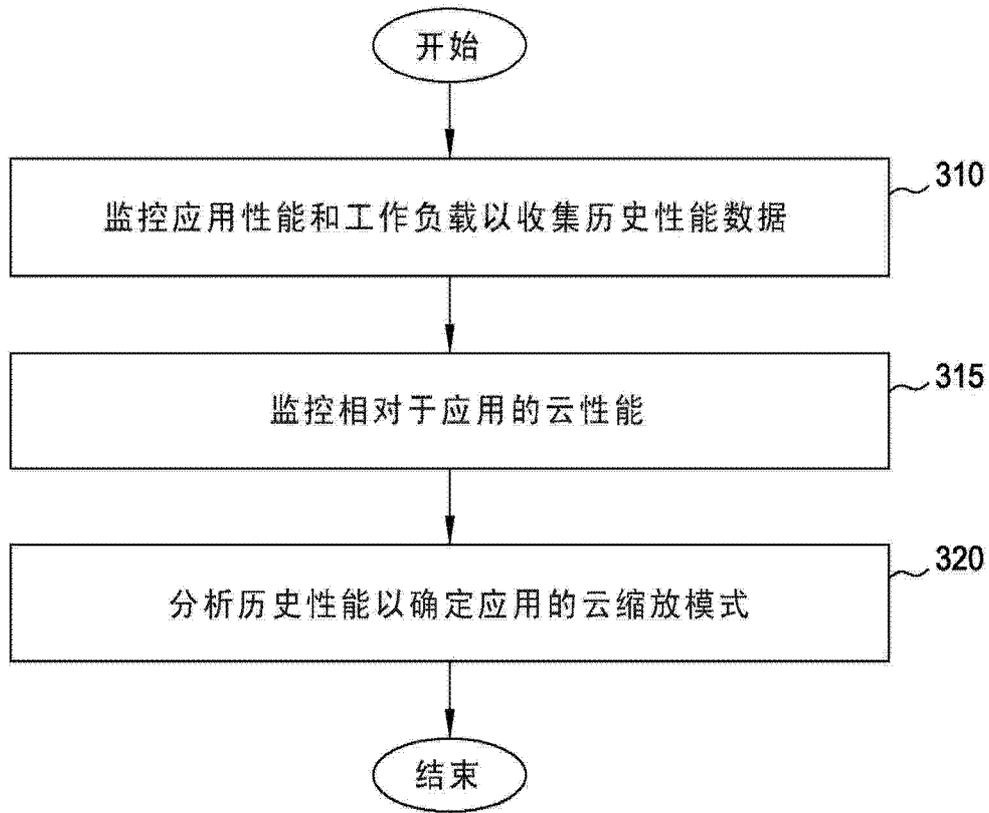


图 3

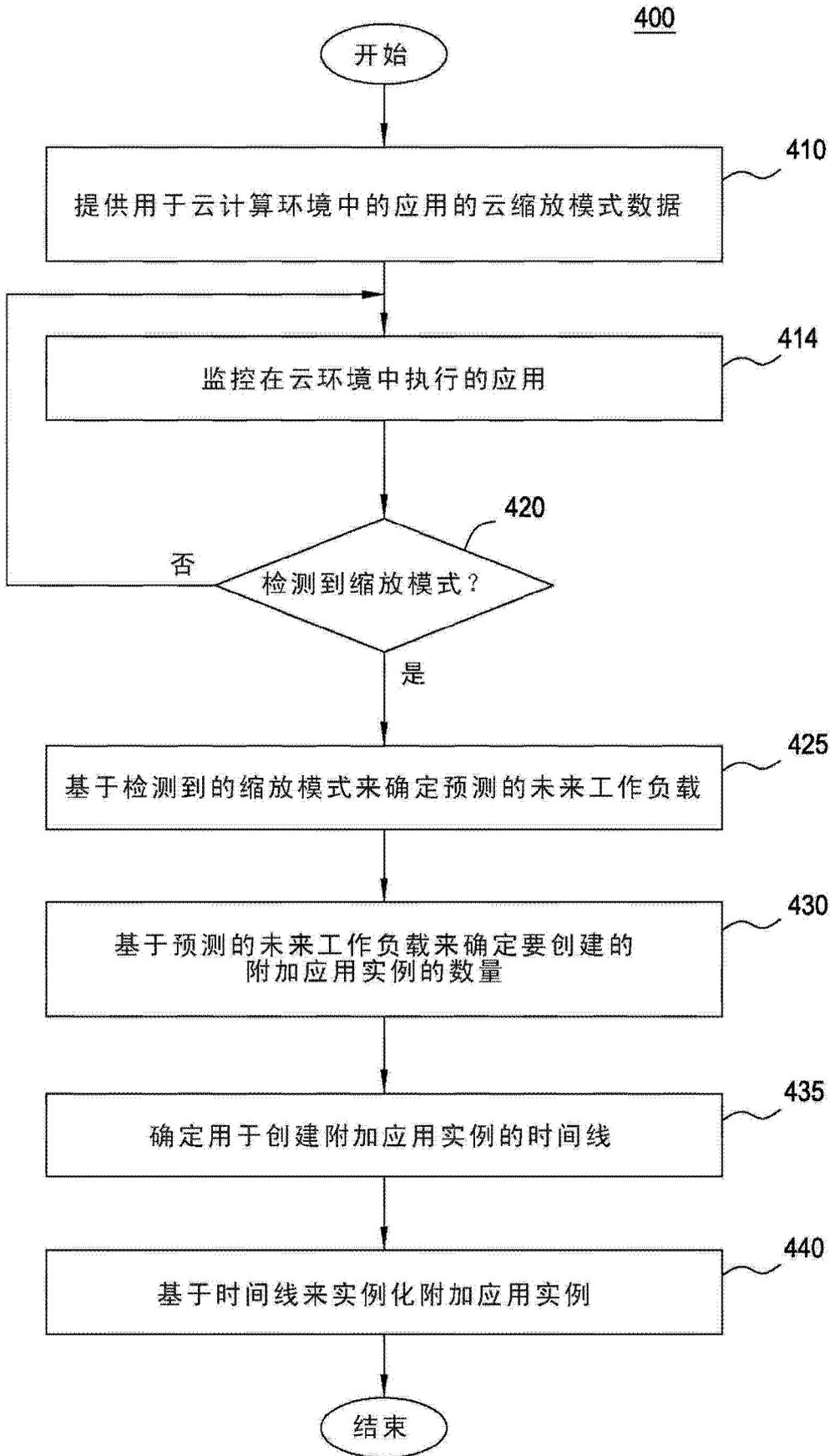


图 4

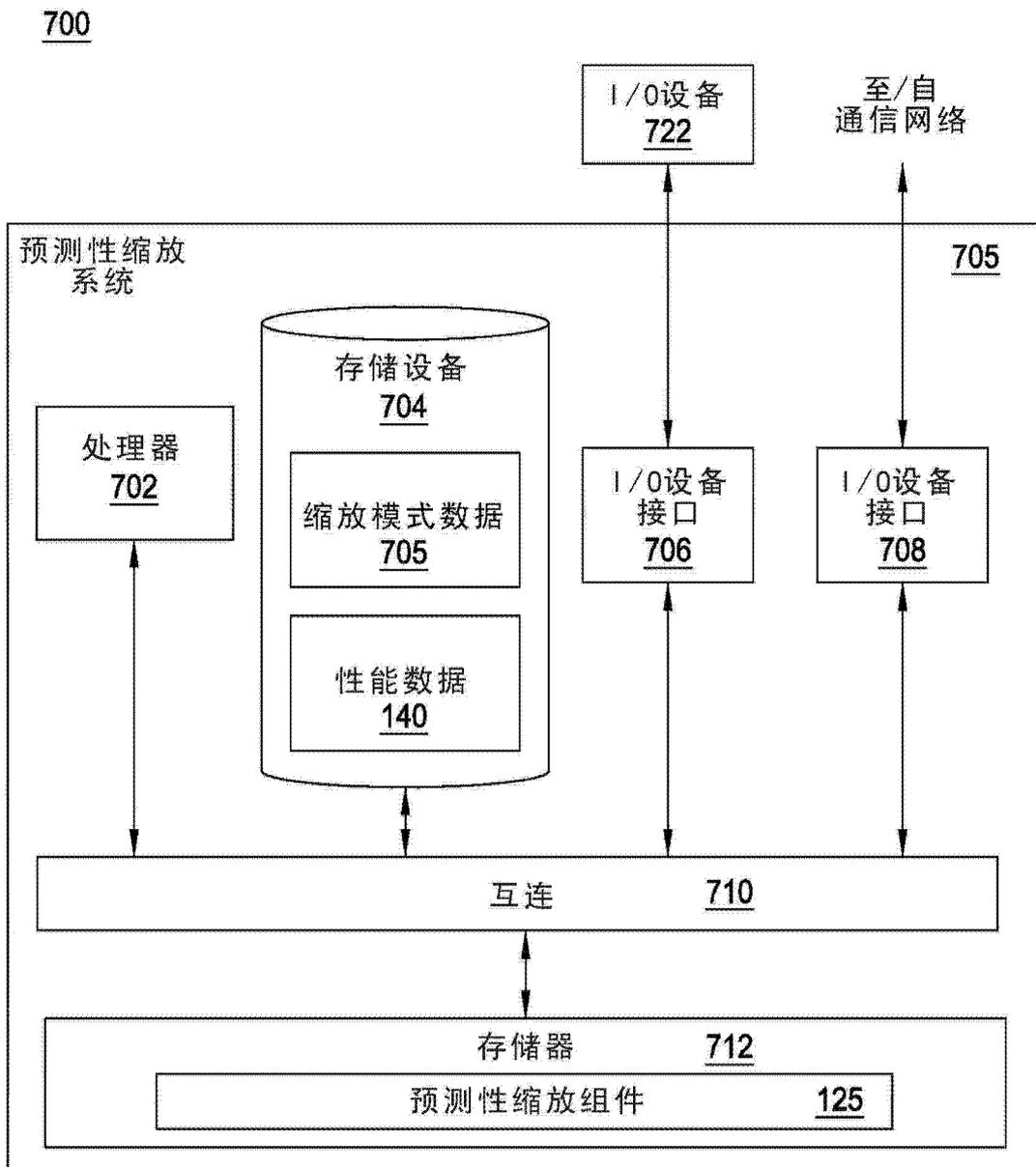


图 5