(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102004671 B
(45)授权公告日 2013.03.13

(21)申请号 201010543864.X

(22)申请日 2010.11.15

(73)专利权人 北京航空航天大学
地址 100191 北京市海淀区学院路 37 号北航计算机学院

(72)发明人 祝明发 王海燕 张振中 肖利民 阮利

(74)专利代理机构 北京慧泉知识产权代理有限公司 11232
代理人 王顺荣 唐爱华

(51)Int.Cl.
G06F 9/50(2006.01)

(56)对比文件
CN 101442807 A,2009.05.27,
CN 1538297 A,2004.10.20,
US 2003/0149685 A1,2003.08.07,
CN 101719082 A,2010.06.02,

审查员 程琼

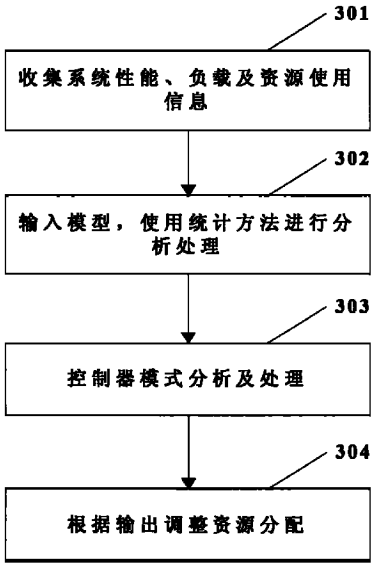
权利要求书 1 页 说明书 5 页 附图 4 页

(54)发明名称

一种云计算环境下数据中心基于统计模型的资源管理方法

(57)摘要

本发明一种云计算环境下数据中心基于统计的资源管理方法机群，该方法有四大步骤：步骤一：收集工作负载、应用程序性能及资源使用状况信息；步骤二：输入模型，使用统计分析方法 KCCA 及远距离相关算法进行模型分析；步骤三：根据当前环境对工作模式进行分类，根据控制参数对资源分配进行调整；步骤四：根据控制器输出调整资源分配，同时更新资源状态。本发明首先考虑云计算环境下新型数据中心工作负载不断变化的特性，根据用户需求及资源使用情况对数据中心资源进行实时监控及弹性管理，以保证在系统性能不受影响的情况下整体资源消耗达到最小。它在云计算数据中心弹性资源管理技术领域里具有广泛地实用价值和应用前景。



CN 102004671 B

1. 一种云计算环境下数据中心基于统计模型的资源管理方法，它采用实时负载及资源监控方式，通过统计模型分析处理监控数据，生成理想分配数据，通过控制器结合实际情况进行处理，输出实际分配数据指导系统资源分配，其特征在于：该方法包括以下步骤：

步骤一：动态收集应用程序负载信息及数据中心资源使用信息；

步骤二：将收集到的信息作为输入数据输入至统计模型，主体采用 KCCA

算法及远距离相关算法对数据进行分析及预测，按照输出调整资源分配；

步骤三：根据模型控制器中设计的四种工作模式，定义及调整控制参数，

使用机器学习的方法调整资源分配数量以满足 SLA；

步骤四：输出分析数据，指导系统进行资源分配，同时更新数据。

2. 根据权利要求 1 所述的一种云计算环境下数据中心基于统计模型的资源管理方法，其特征在于：步骤一所述的应用程序负载信息主要是每秒请求数，即在当前时段使用该应用程序的个数；根据不同的应用程序所需资源的不同，还需要收集 CPU、I/O 及内存的资源使用信息，根据使用状况及剩余资源评估虚拟机的最大容量，以满足应用程序性能需求，避免因资源分配时超过虚拟机最大容量而导致的性能下降，应用程序负载信息及数据中心资源使用信息的收集周期以 0.8 秒～ 1.2 秒之间为最佳。

3. 根据权利要求 1 所述的一种云计算环境下数据中心基于统计模型的资源管理方法，其特征在于：步骤二所述的 KCCA 算法具有处理非线性数据的良好性能，主要根据输入数据预测未来 5 分钟内的负载状况，考虑节点位置、资源使用状况生成应用程序性能及资源模型，进一步计算出理想的资源分配情况，远距离相关算法通过相关参数计算应用程序资源需求间的相关程度，选取最小相关系数，保证分配的资源具有最小相关性，避免资源冲突，实现资源弹性有效供给，统计模型中的数据分析算法由管理员根据系统实际情况做相应的调整。

4. 根据权利要求 1 所述的一种云计算环境下数据中心基于统计模型的资源管理方法，其特征在于：步骤三所述的工作模式是根据现有的 Web 应用程序负载变化规律，分为昼夜模式、工作日 / 周末模式、重大假日模式及用户自主模式四种模式；控制器反映了被控变量即当前测量输出的性能参数和被处理变量即过去和当前的参考输入，分配的各种资源之间的动态关系；不同的工作模式下所需要的应用程序负载信息及数据中心资源使用信息的收集周期以及由于控制器震荡所引起的控制参数都会有所不同，从而优化虚拟资源和应用服务之间的映射关系，满足应用服务的应用服务级别目标 SLO；上述工作模式由管理员根据系统实际情况做相应的调整。

5. 根据权利要求 1 所述的一种云计算环境下数据中心基于统计模型的资源管理方法，其特征在于：步骤四所述的分析数据是指通过统计模型及控制器策略后所得到的资源分配数据；该数据通过统计模型分析后加入控制参数，结合当前的实际情况进行分析，避免因系统情况估计不足所引发的资源冲突或有些应用无法获得足够的资源这种情况的发生，资源分配完成之后，更新剩余资源数据，实现一轮资源分配流程。

一种云计算环境下数据中心基于统计模型的资源管理方法

（一）技术领域

[0001] 本发明涉及一种云计算环境下新型数据中心采用统计学习模型及机器学习的资源管理方法，尤其涉及一种云计算环境下数据中心基于统计模型的资源管理方法，它在线管理负载及资源分配方面，提升系统性能，属于云计算弹性资源管理领域。

（二）背景技术

[0002] 目前，随着网络应用的飞速发展使得对计算能力的需求不断增加，现有的数据中心服务器数量不断增加，如何管理庞大的服务器集群成为大家关注的焦点。随着云计算的流行，越来越多的网络 (Web) 服务和商业应用被部署到云计算环境中，对于云计算数据中心中海量的服务器，如何有效管理云计算数据中心的资源，提升资源利用率同时使用成本达到最低成为当前云计算领域研究的热点。

[0003] 除了从使用基准测试程序 (benchmark) 进行实验仿真外，通过建立资源及性能间的关系模型并通过模型对资源进行管理是目前常用的方法之一，即利用小型 benchmark 或部分服务器收集到的数据进行统计分析以此来衡量资源及性能间的关系，来对数据中心资源进行管理。这种方式下收集到的数据并不能代表生产运行环境中的整体数据，同时无法处理云计算环境下不断变化的工作负载。此方式对资源的管理不灵活，不能根据数据中心用户的实际需求，对资源及性能进行动态调整，有可能会在发生工作负载激增的情况下，系统性能大幅度下降。

[0004] 如果采用动态控制技术，通过在线收集策略及统计模型，根据工作负载的变化自动对资源进行管理分配，可以避免上述静态管理所带来的缺点，并且可以是使系统满足弹性资源管理的需求，在云计算新型数据中心领域具有很重要的意义。

（三）发明内容

[0005] 1、目的：有鉴于此，本发明的目的是提供一种云计算环境下数据中心基于统计模型的资源管理方法，它首先收集应用程序性能参数及资源使用状况，在满足资源需求及性能 SLA 的情况下对数据中心应用程序性能进行动态控制，从而达到使用最少的资源达到最大的性能这样一个目标。

[0006] 为实现上述目的，本发明提出了云计算新型数据中心弹性资源管理实施方法，数据中心结构如图 1 所示。包括：主控节点通过网络适配器与数据网络互联，监控及控制多个节点及虚拟机，主控制器负责对下游借点控制器进行统一调配管理，包括添加、删除及迁移控制系统的任何数量的可读数据的物理驱动器和存储介质等操作，管理模型负责对收集到的负载及资源使用信息进行分析处理，然后交由控制器进行控制。计算节点包含任意数量的虚拟机，每一个节点内部包含一个节点控制器负责节点内部的虚拟机资源控制，资源驱动器负责资源的分配管理，多个包含应用程序资源使用监控器及性能监控器在内的虚拟机以及虚拟机监控器。

[0007] 2、技术方案：为达到上述目的，本发明的技术方案是这样的：

[0008] 如图 2 所示，一种云计算环境下数据中心基于统计模型的资源管理方法，它采用实时负载及资源监控方式，通过统计模型分析处理监控数据，生成理想分配数据，通过控制器结合实际情况进行处理，输出实际分配数据指导系统资源分配，其特征在于：该方法包括以下步骤：

[0009] 步骤一：动态收集应用程序负载信息及数据中心资源使用信息；

[0010] 步骤二：将收集到的信息作为输入数据输入至统计模型，主体采用 KCCA 算法及远

[0011] 距离相关算法对数据进行分析及预测，按照输出调整资源分配；

[0012] 步骤三：根据模型控制器中设计的四种工作模式，定义及调整控制参数，使用机器学习的方法调整资源分配数量以满足 SLA；

[0013] 步骤四：输出分析数据，指导系统进行资源分配，同时更新数据。

[0014] 其中，步骤一所述的应用程序负载信息主要是每秒请求数，即在当前时段使用该应用程序的个数；根据不同的应用程序所需资源的不同，还需要收集 CPU、I/O 及内存的资源使用信息，根据使用状况及剩余资源评估虚拟机的最大容量，以满足应用程序性能需求，避免因资源分配时超过虚拟机最大容量而导致的性能下降，应用程序负载信息及数据中心资源使用信息的收集周期以 0.8 秒～ 1.2 秒之间为最佳。

[0015] 其中，步骤二所述的 KCCA 算法具有处理非线性数据的良好性能，主要根据输入数据预测未来 5 分钟内的负载状况，考虑节点位置、资源使用状况生成应用程序性能及资源模型，进一步计算出理想的资源分配情况，远距离相关算法通过相关参数计算应用程序资源需求间的相关程度，选取最小相关系数，保证分配的资源具有最小相关性，避免资源冲突，实现资源弹性有效供给，统计模型中的数据分析算法由管理员根据系统实际情况做相应的调整。

[0016] 其中，步骤三所述的工作模式是根据现有的 Web 应用程序负载变化规律，分为昼夜模式、工作日 / 周末模式、重大假日模式及用户自主模式四种模式；控制器反映了被控变量即当前测量输出的性能参数和被处理变量即过去和当前的参考输入，分配的各种资源之间的动态关系；不同的工作模式下所需要的应用程序负载信息及数据中心资源使用信息的收集周期以及由于控制器震荡所引起的控制参数都会有所不同，从而优化虚拟资源和应用服务之间的映射关系，满足应用服务的应用服务级别目标 SLO；上述工作模式由管理员根据系统实际情况做相应的调整。

[0017] 其中，步骤四所述的分析数据是指通过统计模型及控制器策略后所得到的资源分配数据；该数据通过统计模型分析后加入控制参数，结合当前的实际情况进行分析，避免因系统情况估计不足所引发的资源冲突或有些应用无法获得足够的资源这种情况的发生，资源分配完成之后，更新剩余资源数据，实现一轮资源分配流程。

[0018] 3、优点及功效：本发明一种云计算环境下数据中心基于统计模型的资源管理方法，它与现有技术比，其主要优点是：(1) 提升原有的数据中心资源供应效率，节省开销。原有的数据中心为避免违反服务水平协议 (SLA) 而向用户支付赔偿金而采取过量供应机制，为数据中心带来很大的开销。本方法根据数据中心应用负载及性能进行动态调整，通过控制理论进一步保证了资源分配的准确性，确保了应用程序的性能去求；(2) 将统计机器学习方法和控制理论到虚拟化计算系统的资源管理中，构建新的架构、模型和方法以解决计算资源针对应用负载和系统环境变化的自适应问题。(3) 提出了控制器的典型控制模式，针

对不同的模式使用不同的控制调节参数，实现使用最少的资源达到最优的性能这一目的。

（四）附图说明

[0019] 图 1 本发明的云计算环境下新型数据中心结构模型示意图

[0020] 图 2 基于统计机器学习的模型结构示意图

[0021] 图 3 基于模型的资源管理流程示意图

[0022] 图 4 控制器控制资源分配示意图

[0023] 图 5 根据系统负载、性能及资源使用状况进行动态资源分配的流程示意图

[0024] 图 6 统计模型分析模块示意图

[0025] 其中图 4 中符号说明如下：

[0026] 参考输入 (x)：代表分配给虚拟机的资源，如 CPU、内存、网络 I/O 等

[0027] 测量输出 (y)：代表系统希望达到的性能参数，如吞吐量、响应时间等 SLA 中所规定的参数指标

[0028] 控制误差 (e)：是控制精度（准确度）的一种度量，是控制系统的稳态性能指标。它依赖于参考输入 (x) 和测量输出 (y)

[0029] 滞后参数（α，β）：用来决定系统增加或删除资源的速度，其具体数值同上述四种模式相关，α 值越大，表明添加资源的速度越快；β 值越大，表明移除资源的速度越快。

（五）具体实施方式

[0030] 为使本发明的目的、技术方案和优点表达得更加清楚明白，下面结合附图及具体实施例对本发明再作进一步详细的说明。

[0031] 本发明在硬件条件方面，要求各结点同时支持共享存储。在软件条件方面，若操作系统采用的是 Linux，要求其内核版本在 2.6.18 以上，以避免低版本内核在电源管理方面的缺陷。

[0032] 本发明所需满足的设备条件见图 1，该数据中心基础架构包括：主控节点通过网络适配器与数据网络互联，监控及控制多个节点及节点中的多个虚拟机，主控制器负责对下游节点控制器进行统一调配管理，包括添加、删除及迁移控制系统的任何数量的可读数据的物理驱动器和存储介质等操作，管理模型负责对收集到的负载及资源使用信息进行分析处理，然后交由控制器进行控制。计算节点包含任意数量的虚拟机，每一个虚拟机上安装一种应用程序（例如 Web2.0），每一个节点内部包含一个节点控制器负责节点内部的虚拟机资源控制，资源驱动器负责资源的分配管理，多个包含应用程序资源使用监控器及性能监控器在内的虚拟机以及虚拟机监控器。

[0033] 基于统计的资源管理模型见图 2，本发明采用实时负载及资源监控方式，通过统计模型分析处理监控数据，生成理想分配数据，通过控制器结合实际情况进行处理，输出实际分配数据指导系统资源分配，以保证在使用较少的资源情况下实现系统最大的性能，同时节省系统开销。

[0034] 数据中心收集到的信息包括：数据中心相应的应用程序负载信息，以每秒请求数为例，分析用户对于该应用程序的需求；收集用户服务水平协议 SLA 中所涉及到的用户需求指标，如吞吐率及响应时间等，以此作为系统性能衡量指标；数据中心资源使用情况，如

CPU 使用率、I/O 及网络带宽等数据，用来衡量系统使用状况及系统整体容量。需要实时收集上述信息，及时更新数据，保证数据中心服务质量。

[0035] 下面以一实例进行说明，如图 5 所示，包括以下步骤：

[0036] 步骤 501：实时收集应用程序负载、性能及资源使用信息系统。同时记录所获取的信息，更新周期设为 1.0 秒。

[0037] 步骤 502：将上述步骤所收集到的信息作为输入，输入到统计模型模块中进行分析，主要分析应用程序负载、性能及资源使用间的关系，将其输出作为步骤 503 的输入。具体如图 6 统计模型分析模块所示。

[0038] 步骤 503：控制器模式分析。根据用户所需要的虚拟机资源，如 CPU、内存等，利用上述统计模型给出的资源分配方案，考虑外界环境对系统产生的干扰（如系统的管理和维护需求），根据当前分析的工作模式，采取不同的控制模式。在不同的模式下，控制器滞后参数是不同的，例如在重大假日模式中，设置 $\alpha = 0.9$ 可以对系统异常等进行快速反应，快速加入资源，$\beta = 0.01$ 保证系统在需要移除资源时可以更为保守的估计，不需要快速反应。该组参数适用于需要快速反应的模式。在其他模式下，适当调整该组参数，使其系统对工作负载的变化反应满足用户需求同时不需要消耗过多的资源。

[0039] 步骤 504：根据控制器输出的参数控制资源的分配。首先由主控制器分配任务给节点控制器，节点控制器根据具体应用需求分配满足要求的虚拟机。

[0040] 其中统计模型分析模块流程如图 6 所示，包括以下步骤：

[0041] 步骤 601：在系统运行中，根据图 5 步骤 501 所收集到的信息，实时监控系统参数生成的曲线，观察曲线是否有突变点或不符合拟合曲线的异常值出现。若存在，则说明当前数据中心该应用出现尖峰时刻，则应采取步骤 602 进行处理；若不存在，则说明系统运行稳定，则进行步骤 603。

[0042] 步骤 602：系统曲线出现异常值点，证明系统所需资源需要进行较大的变动。此时需要快速对系统资源使用状况进行分析，当系统资源达到最大容量时是否可以满足资源需求。因为是异常值点，当前所获参数不能代表整体的负载、性能及资源间的关系模型，但如果不及时处理的话会对系统的性能有很大的影响，所以需要及时对异常值点进行分析处理。若系统资源池资源满足当前需求，则直接交给控制器进行处理，快速解决当前异常值点；若当前系统资源达到最大容量时也无法满足应用程序性能需求，需要及时采取虚拟化技术，如虚拟机迁移等快速加入资源，满足用户资源需求。即按需供应，节省系统资源开销。将加入的资源放入系统资源池，通过控制器统一控制。

[0043] 步骤 603：使用简单的线性回归模型预测下一个 5 分钟的工作负载，即每秒应用程序请求数。简单的线性回归模型可以有效的捕捉工作负载随时间变化规律，即使是更为复杂的历史数据也可以很容易的归纳预测其负载。

[0044] 步骤 604：预测的工作负载作为模型的输入来评估现有的工作量所需的资源需求及系统可以达到的性能。许多复杂的因素都会影响应用程序的性能，例如混合负载、应用程序代码的改变等等，采用 KCCA 算法及远距离相关算法实现多元统计分析建模，同时分析多个影响因素对系统性能带来的影响，实时调整模型参数，生成理想资源分配数据。最后，将其输出作为图 5 步骤 503 的输入，跳转到控制器模式分析。

[0045] 图 3 是本发明基于模型的资源管理流程示意图；图 4 是控制器控制资源分配示意

图。

[0046] 本实例中查看资源使用及工作负载等参数并进行相应的参数更新时在资源分配过程中依据资源采集周期循环执行的。采用上述弹性资源管理方法可以做到尽量保证在任何时刻系统使用较少的资源达到最大的性能，以满足用户的需求。

[0047] 最后所应说明的是：以上实施例仅用以说明而非限制本发明的技术方案，尽管参照上述实施例对本发明进行了详细说明，本领域的普通技术人员应当理解：依然可以对本发明进行修改或者等同替换，而不脱离本发明的精神和范围的任何修改或局部替换，其均应涵盖在本发明的权利要求范围当中。

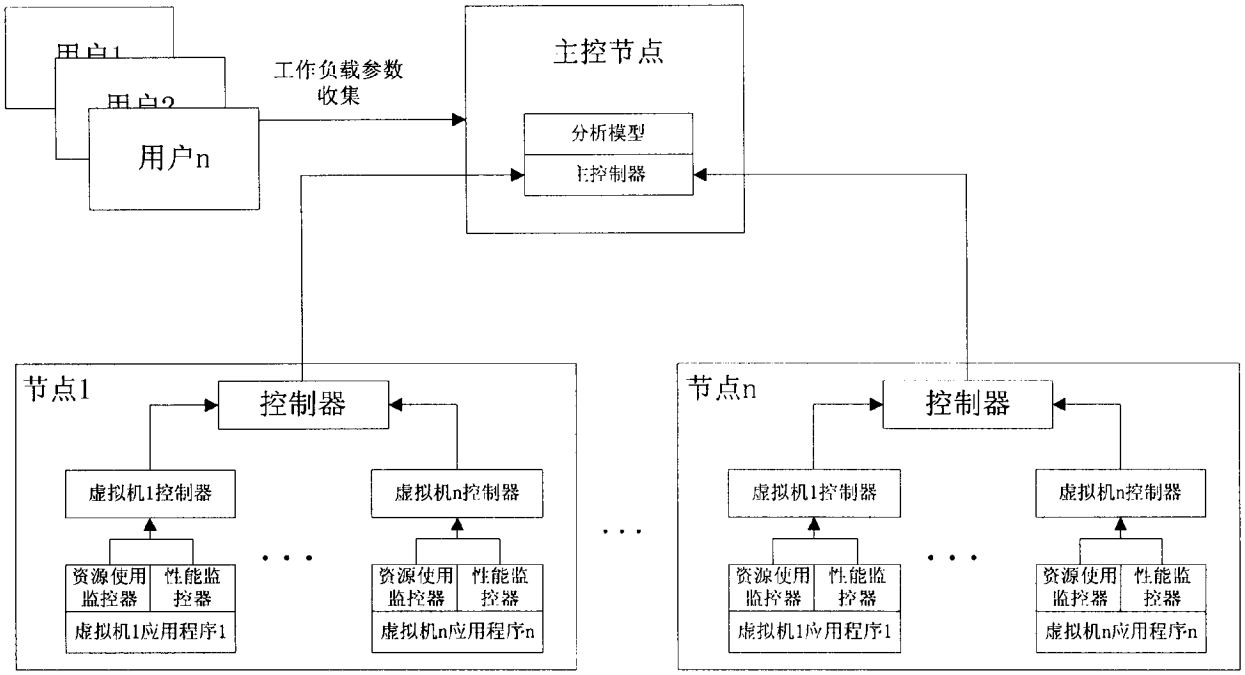

用户1
用户2
用户n
工作负载参数
收集
主控节点
分析模型
主控制器
节点1
控制器
虚拟机1控制器
虚拟机n控制器
资源使用监控器
性能监控器
虚拟机1应用程序1
虚拟机n应用程序n
节点n
控制器
虚拟机1控制器
虚拟机n控制器
资源使用监控器
性能监控器
虚拟机1应用程序1
虚拟机n应用程序n

图 1

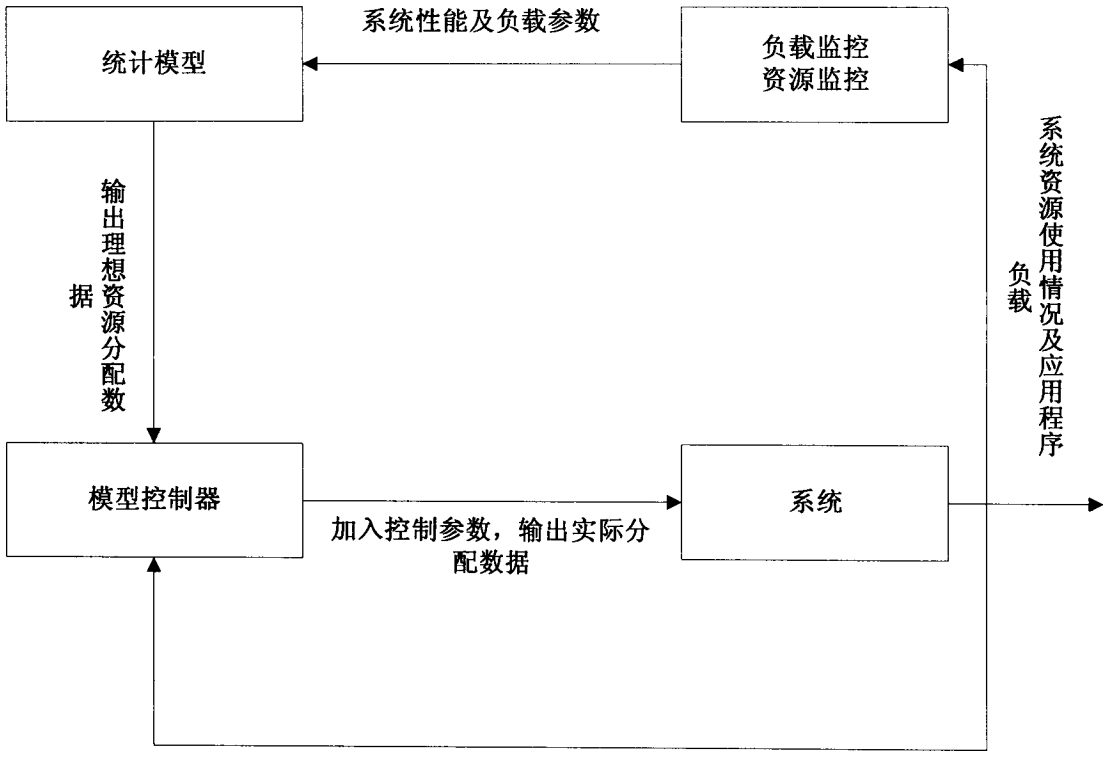

统计模型
系统性能及负载参数
负载监控
资源监控
输出理想资源分配数据
系统资源使用情况及应用程序负载
模型控制器
加入控制参数，输出实际分配数据
系统

图 2

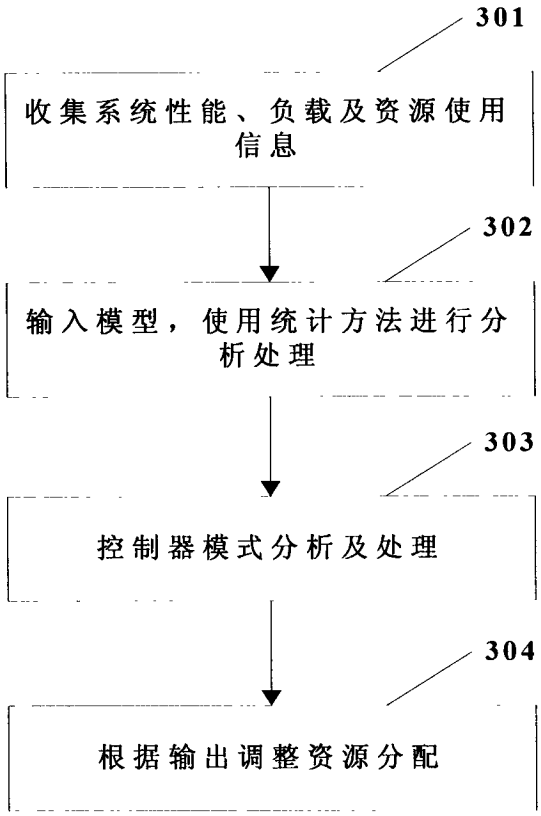

301
收集系统性能、负载及资源使用信息
302
输入模型，使用统计方法进行分析处理
303
控制器模式分析及处理
304
根据输出调整资源分配

图 3

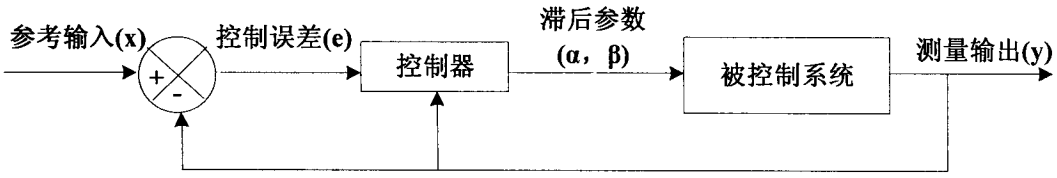

参考输入(x)
控制误差(e)
+
-
控制器
滞后参数
(α，β)
被控制系统
测量输出(y)

图 4

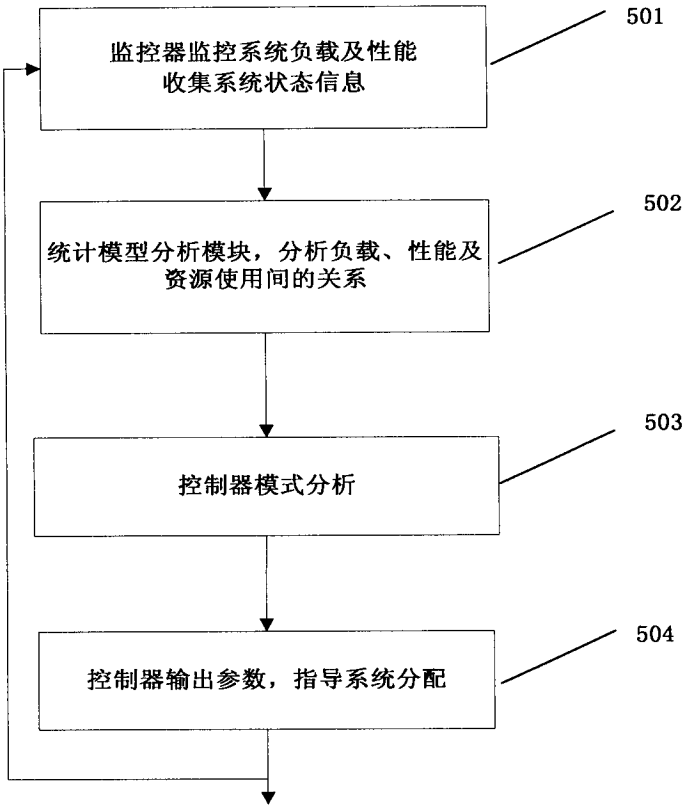

监控器监控系统负载及性能
收集系统状态信息
501
统计模型分析模块，分析负载、性能及
资源使用间的关系
502
控制器模式分析
503
控制器输出参数，指导系统分配
504

图 5

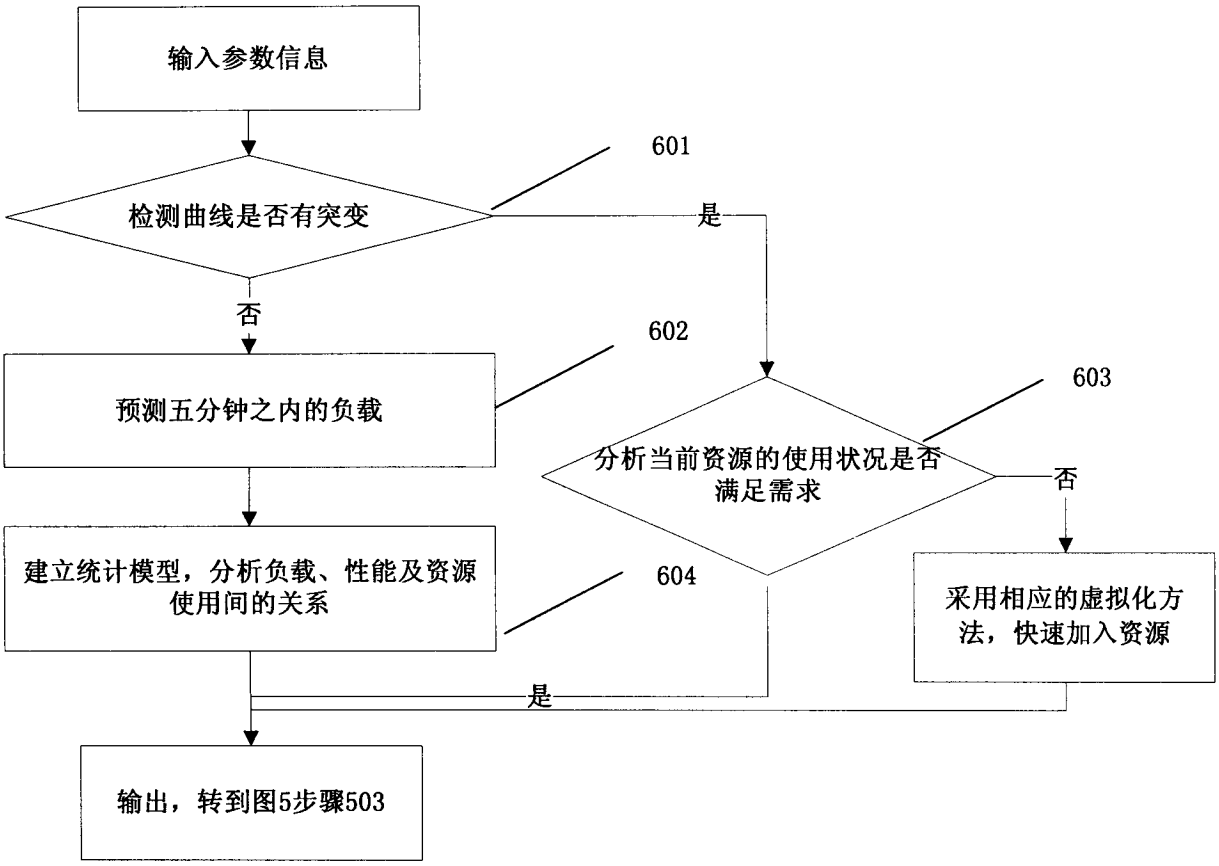

输入参数信息
601
检测曲线是否有突变
是
否
602
预测五分钟之内的负载
603
分析当前资源的使用状况是否满足需求
否
建立统计模型，分析负载、性能及资源使用间的关系
604
采用相应的虚拟化方法，快速加入资源
是
输出，转到图5步骤503

图 6