(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102646062 A
(43)申请公布日 2012.08.22

(21)申请号 201210075071.9

(22)申请日 2012.03.20

(71)申请人 广东电子工业研究院有限公司
地址 523808 广东省东莞市松山湖科技产业园区松科苑 10 号楼

(72)发明人 王志荣 岳强 季统凯

(74)专利代理机构 北京科亿知识产权代理事务所(普通合伙) 11350
代理人 汤东凤

(51)Int.Cl.
G06F 9/50(2006.01)
G06F 9/455(2006.01)
H04L 29/08(2006.01)

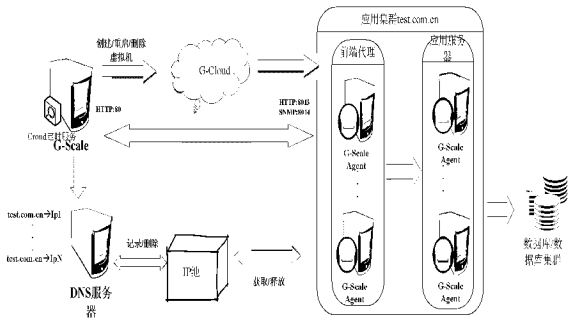
权利要求书 2 页 说明书 4 页 附图 2 页

(54)发明名称

一种云计算平台应用集群弹性扩容方法

(57)摘要

本发明涉及云计算技术领域,特指一种基于云计算平台应用集群弹性扩容方法。本发明根据应用集群中虚拟机的负载压力变化,弹性调整集群中虚拟机的规模,实现对云计算平台各种资源的有效地利用。本发明自动搭建应用程序集群,智能化管理集群的规模以有效应对负载压力的变化,使云平台资源得到最优化利用;实现了应用集群部署自动化、扩容弹性化、智能化。本发明可应用于云计算平台应用集群的扩容中。



CN 102646062 A

1. 一种云计算平台的应用集群弹性扩容方法，其特征在于：根据应用集群中虚拟机的负载压力变化，弹性调整集群中虚拟机的规模，实现对云计算平台各种资源的有效地利用。

2. 根据权利要求 1 所述的应用集群弹性扩容方法，其特征在于：所述的方法包括：

虚拟机管理控制机制，通过云计算平台提供的 web service 接口，实现对虚拟机的创建、删除、停止、启动等控制虚拟机本身的操作；

基于角色的消息事件机制，将应用集群中的虚拟机群以功能用途划分为不同的角色，一般为前端代理和应用服务器；角色具体体现在虚拟机上安装的不同功能软件，如前端代理角色的虚拟机安装 nginx，而应用服务器上可能安装 apche 或 IIS；每种角色除了拥有 HostInit、HostUp、HostDown 等公共事件外，还拥有自身特有的事件；

基于 SNMP 协议的虚拟机监控机制，部署在虚拟机上的 snmp agent 收集 CPU、内存、网络等数据，而控制端通过 SNMP 协议获取所需的实时监控数据；控制端的实现可以采用现成的大量 SNMP Manager 工具，以减少开发的工作量；

可扩展的调度策略机制，根据监控数据对应用集群中的虚拟机进行弹性扩容，具体形式是增加 / 减少某角色下虚拟机的数量或提高 / 降低虚拟机的配置；简单的策略可以只对虚拟机数量增减，复杂的策略可以是当集群规模的增大、网络通讯影响性能时，提高虚拟机配置，减少虚拟机数量；

在前述机制基础上，由虚拟机管理控制机制创建应用集群中不同角色的服务器，基于角色的消息事件机制自动部署形成应用集群的配置，并在不同时间段控制虚拟机上的功能软件，基于 SNMP 协议的虚拟机监控机制提供应用集群弹性扩容的数据依据，可扩展的调度策略机制则根据实际情况对应用集群进行扩容操作。

3. 根据权利要求 1 或 2 所述的应用集群弹性扩容方法，其特征在于：所述的方法涉及云计算平台、G-Scale、和 G-Scale Agent；

所述的云计算平台为由国云科技有限公司自主研发的 IAAS 云计算平台，由其提供计算资源，网络资源和存储资源，以 REST 风格的 Web Service 接口供 G-Scale 使用；

所述的 G-Scale 是一个云计算平台的管理调度工具，通过调用 G-Cloud 接口，使用预先制作好的镜像创建虚拟机；当虚拟机成功启动后，G-Scale 的后台服务会自动将虚拟机群搭建成应用集群；G-Scale 会开放 80 端口，为虚拟机上的 G-Scale Agent 进程提供消息事件服务；同时 G-Scale 也会主动向 Scale Agent 发送请求，获取各虚拟机有用信息，向 Agent 的 8013 端口发送消息请求，通过 SNMP 协议在 8014 端口获取虚拟机的 CPU、内存、带宽等负载压力信息；G-Scale 的后台定时服务会通过 SNMP 协议获取各虚拟机的负载压力信息，根据预先设定好的调度策略，增加 / 减少虚拟机数量，提高 / 降低虚拟机配置，实现弹性扩容；

所述的 G-Scale Agent 是一个随虚拟机启动而自动运行的进程，实现与 Scale 进行消息事件交互和提供基于 SNMP 协议的监控服务；Agent 主动向 Scale 的 80 端口发送消息信息，并在 8013 端口监听 Scale 的消息请求；在 8014 端口为 Scale 提供监控虚拟机负载压力重要信息的服务。

4. 根据权利要求 3 所述的应用集群弹性扩容方法，其特征在于：具体包括以下步骤：

第 1 步，通过 G-Scale 提供的 WEB 控制台，创建应用集群，设定好前端代理与应用服务器角色，并为两种角色设定好扩容策略，包括一个由最小最大值组成的范围段；

第 2 步，G-Scale 后台定时服务调用 G-Cloud 的创建虚拟机接口，为每种角色创建虚拟机；虚拟机成功启动后，经过 HostInit、HostUp 等若干个消息事件交互后，建立应用集群；

第 3 步，G-Scale 另一后台服务定时向应用集群中的各虚拟机发送 SNMP 请求，收集各虚拟机的负载压力数据，并根据第 1 步中设定的扩容策略，如果压力数据的平均值或总和处于策略设定的范围段内，G-Scale 不作任何处理，否则 G-Scale 就会作出增加 / 减少虚拟机数量或提高 / 降低虚拟机配置‘另外如果应用集群中虚拟机数量达到一定规模，由于应用集群内的虚拟机网络通讯随着虚拟机数量规模的增大会消耗不少系统资源，G-Scale 此时会减少虚拟机数量，增强部分虚拟机的性能配置，使应用集群发挥更好的性能。

一种云计算平台应用集群弹性扩容方法

[0001] 发明名称

[0002] 本发明涉及云计算技术领域，特指一种基于云计算平台应用集群弹性扩容方法。

背景技术

[0003] IAAS 云计算平台 :IAAS 整合分布式计算、网络计算、虚拟化等技术，提供统一的计算、网络、存储的资源，以虚拟机的形式提供给用户使用。对应用程序而言，以往部署在物理机上，现在直接由云平台统一管理，以虚拟机的形式提供应用程序部署环境。

[0004] 负载均衡（集群）:应用程序部署到单台服务器上，随着用户量的不断增加，负载压力也不断增加。提高服务器的配置是解决问题的一种方法，但单台服务器的服务能力总会有极限 ;此时增加服务器数量是一种节约成本，行之有效的解决方法。负载均衡技术就是这种方法。通过前端的负载均衡器，将后端的应用服务器组成集群，通过增加 / 减少机器的数量达到应用伸缩的效果，即使面对百万，及至千万级的用户都不成问题。

[0005] 目前云计算技术的一个好处可以有效的解决服务器利用率低的问题，对应用程序不需要再以物理服务器为单位分配资源，在需要时可以直接创建虚拟机为应用程序提供部署，在不需要时可以删除虚拟机，释放服务器资源。而负载均衡技术可以有效解决大用户量应用程序苛刻的运行环境。

[0006] 无论部署环境是物理机还是虚拟机，负载均衡技术都有一定的部署难度，自动化程度不高，更多需要系统管理员一步一步地按需求进行配置，部署。应用用户量大时，需要手动增加服务器 ;用户量降低后，同样需要手动撤掉某台服务器，再将它分配另外的应用。很难做到自动化、智能化。

发明内容

[0007] 本发明解决的技术问题是针对云平台上部署应用程序集群过于复杂麻烦，不能有效方便利用云平台计算、网络、存储资源的问题，提出一种云计算平台的应用集群弹性扩容方法。

[0008] 本发明解决上述技术问题的技术方案是 :根据应用集群中虚拟机的负载压力变化，弹性调整集群中虚拟机的规模，实现对云计算平台各种资源的有效地利用。

[0009] 所述的方法包括 :

[0010] 虚拟机管理控制机制，通过云计算平台提供的web service接口，实现对虚拟机的创建、删除、停止、启动等控制虚拟机本身的操作 ;

[0011] 基于角色的消息事件机制，将应用集群中的虚拟机群以功能用途划分为不同的角色，一般为前端代理和应用服务器 ;角色具体体现在虚拟机上安装的不同功能软件，如前端代理角色的虚拟机安装nginx(一个高性能的HTTP和反向代理服务器)，而应用服务器上可能安装 apche(Web 服务器软件）或 IIS(由微软公司提供的基于运行 Microsoft Windows 的互联网基本服务）;每种角色除了拥有 HostInit、HostUp、HostDown 等公共事件外，还拥有自身特有的事件 ;

[0012] 基于SNMP协议的虚拟机监控机制，部署在虚拟机上的snmp agent收集CPU、内存、网络等数据，而控制端通过SNMP协议获取所需的实时监控数据；控制端的实现可以采用现成的大量SNMP Manager工具，以减少开发的工作量；

[0013] 可扩展的调度策略机制，根据监控数据对应用集群中的虚拟机进行弹性扩容，具体形式是增加/减少某角色下虚拟机的数量或提高/降低虚拟机的配置；简单的策略可以只对虚拟机数量增减，复杂的策略可以是当集群规模的增大、网络通讯影响性能时，提高虚拟机配置，减少虚拟机数量；

[0014] 在前述机制基础上，由虚拟机管理控制机制创建应用集群中不同角色的服务器，基于角色的消息事件机制自动部署形成应用集群的配置，并在不同时间段控制虚拟机上的功能软件，基于SNMP协议的虚拟机监控机制提供应用集群弹性扩容的数据依据，可扩展的调度策略机制则根据实际情况对应用集群进行扩容操作。

[0015] 所述的方法涉及云计算平台、G-Scale、和G-Scale Agent；

[0016] 所述的云计算平台为由国云科技有限公司自主研发的IAAS云计算平台，由其提供计算资源，网络资源和存储资源，以REST风格的Web Service接口供G-Scale使用；

[0017] 所述的G-Scale是一个云计算平台的管理调度工具，通过调用G-Cloud接口，使用预先制作好的镜像创建虚拟机；当虚拟机成功启动后，G-Scale的后台服务会自动将虚拟机群搭建成应用集群；G-Scale会开放80端口，为虚拟机上的G-Scale Agent进程提供消息事件服务；同时G-Scale也会主动向Scale Agent发送请求，获取各虚拟机有用信息，向Agent的8013端口发送消息请求，通过SNMP协议在8014端口获取虚拟机的CPU、内存、带宽等负载压力信息；G-Scale的后台定时服务会通过SNMP协议获取各虚拟机的负载压力信息，根据预先设定好的调度策略，增加/减少虚拟机数量，提高/降低虚拟机配置，实现弹性扩容；

[0018] 所述的G-Scale Agent是一个随虚拟机启动而自动运行的进程，实现与Scale进行消息事件交互和提供基于SNMP协议的监控服务；Agent主动向Scale的80端口发送消息信息，并在8013端口监听Scale的消息请求；在8014端口为Scale提供监控虚拟机负载压力重要信息的服务。

[0019] 具体包括以下步骤：

[0020] 第1步，通过G-Scale提供的WEB控制台，创建应用集群，设定好前端代理与应用服务器角色，并为两种角色设定好扩容策略，包括一个由最小最大值组成的范围段；

[0021] 第2步，G-Scale后台定时服务调用G-Cloud的创建虚拟机接口，为每种角色创建虚拟机；虚拟机成功启动后，经过HostInit、HostUp等若干个消息事件交互后，建立应用集群；

[0022] 第3步，G-Scale另一后台服务定时向应用集群中的各虚拟机发送SNMP请求，收集各虚拟机的负载压力数据，并根据第1步中设定的扩容策略，如果压力数据的平均值或总和处于策略设定的范围段内，G-Scale不作任何处理，否则G-Scale就会作出增加/减少虚拟机数量或提高/降低虚拟机配置‘另外如果应用集群中虚拟机数量达到一定规模，由于应用集群内的虚拟机网络通讯随着虚拟机数量规模的增大会消耗不少系统资源，G-Scale此时会减少虚拟机数量，增强部分虚拟机的性能配置，使应用集群发挥更好的性能。

[0023] 本发明提出了一种基于G-Cloud云计算平台的应用集群弹性扩容技术，在尽量少

的外界人工干预的情况下，通过程序自动搭建应用程序集群，智能化管理集群的规模以有效应对负载压力的变化，使云平台资源得到最优化利用。本发明实现了应用集群部署自动化、扩容弹性化、智能化。

附图说明

[0024] 下面结合附图对本发明进一步说明：

[0025] 图 1 是本发明云计算平台的应用集群弹性扩容示意图；

[0026] 图 2 是本发明云计算平台的应用集群弹性扩容处理流程图。

具体实施方式

[0027] 见附图 1、2 所示，本发明云计算平台的应用集群弹性扩容方法是根据应用集群中虚拟机的负载压力变化，弹性调整集群中虚拟机的规模，有效地利用 G-Cloud 云计算平台的各种资源。本发明涉及到 4 种机制和 3 个实体。4 种机制分别是：

[0028] 虚拟机管理控制机制：通过云计算平台提供的 web service 接口，实现对虚拟机的创建，删除，停止，启动等控制虚拟机本身的操作。另外云计算平台提供的镜像技术，为应用集群的自动化部署，提供重要技术保障。

[0029] 基于角色的消息事件机制：将应用集群中的虚拟机群以功能用途划分为不同的角色，一般是前端代理和应用服务器。角色的体现在虚拟机上安装的不同功能软件，如前端代理角色的虚拟机安装 nginx，而应用服务器上可能安装 apche 或 IIS。实现对不同角色虚拟机上的功能软件的控制是通过消息事件来完成，每种角色除了拥有 HostInit，HostUp，HostDown 等公共事件外，还会有自身特有的事件。

[0030] 基于 SNMP 协议的虚拟机监控机制：部署在虚拟机上的 snmp agent 收集 CPU，内存，网络等数据，而控制端通过 SNMP 协议可以轻易获取所需的实时监控数据。控制端的实现可以采用现成的大量 SNMP Manager 工具，以减少开发的工作量。

[0031] 可扩展的调度策略机制：根据监控数据对应用集群中的虚拟机进行弹性扩容，具体形式是增加 / 减少某角色下虚拟机的数量或提高 / 降低虚拟机的配置。调试策略的可扩展性使对应用集群的控制随心所欲，简单的策略可以只对虚拟机数量的增减，复杂的策略可以是当集群规模的增大，网络通讯影响性能时，提高虚拟机配置，减少虚拟机数量。

[0032] 4 种机制是这样形成应用集群弹性扩容技术：虚拟机管理控制机制创建应用集群中不同角色的服务器，基于角色的消息事件机制自动部署形成了应用集群的配置，并在不同时间段控制虚拟机上的功能软件，基于 SNMP 协议的虚拟机监控机制提供了应用集群弹性扩容的数据依据，可扩展的调度策略机制则根据实际情况对应用集群进行真正的操作。

[0033] 本发现的 3 个实体分别是：

[0034] G-Cloud：一个由国云科技有限公司自主研发的 IAAS 云计算平台，在本技术中主要提供计算资源，网络资源和存储资源，以 REST 风格的 Web Service 接口供 G-Scale 使用。

[0035] G-Scale：一个云计算平台的管理调度工具，通过调用 G-Cloud 接口，使用预先制作好的镜像创建虚拟机（其中预先制作好的镜像包括前端代理专用和应用服务器专用两种不同的镜像），当虚拟机成功启动后，G-Scale 的后台服务会自动将虚拟机群搭建成应用集群；G-Scale 会开放 80 端口，为虚拟机上的 G-ScaleAgent 进程提供消息事件服务；同时

G-Scale 也会主动向 Scale Agent 发送请求，获取各虚拟机有用信息，向 Agent 的 8013 端口发送消息请求，8014 端口通过 SNMP 协议获取虚拟机的 CPU，内存，带宽等负载压力信息。G-Scale 的后台定时服务会通过 SNMP 协议获取各虚拟机的负载压力信息，根据预先设定好的调度策略，增加 / 减少虚拟机数量，提高 / 降低虚拟机配置，实现弹性扩容。

[0036] G-Scale Agent：一个随虚拟机启动而自动运行的进程，主要有两大作用：与 Scale 进行消息事件交互和提供基于 SNMP 协议的监控服务。Agent 主动向 Scale 的 80 端口发送消息信息，并在 8013 端口监听 Scale 的消息请求；在 8014 端口为 Scale 提供监控虚拟机负载压力重要信息的服务。

[0037] 本发明实现了基于 G-Cloud 云计算平台的应用集群弹性扩容方法，如图 2 所示，方法具体步骤如下：

[0038] 第 1 步，通过 G-Scale 提供的 WEB 控制台，创建应用集群，设定好前端代理与应用服务器角色，并为两种角色设定好扩容策略（包括一个由最小最大值组成的范围段）。

[0039] 第 2 步，G-Scale 后台定时服务会调用 G-Cloud 的创建虚拟机接口，为每种角色（前端代理和应用服务器）创建虚拟机。虚拟机成功启动后，经过 HostInit，HostUp 等若干个消息事件交互后，应用集群会成功建立。

[0040] 第 3 步，G-Scale 另一后台服务会定时向应用集群中的各虚拟机发送 SNMP 请求，收集各虚拟机的负载压力数据，并根据第 1 步中设定的扩容策略，如果压力数据的平均值或总和处于策略设定的范围段内，G-Scale 不作任何处理，否则 G-Scale 就会作出增加 / 减少虚拟机数量或提高 / 降低虚拟机配置。另外如果应用集群中虚拟机数量达到一定规模，由于应用集群内的虚拟机网络通讯随着虚拟机数量规模的增大而消耗不少系统资源，所以 G-Scale 此时会减少虚拟机数量，增强部分虚拟机的性能配置，使应用集群发挥更好的性能。

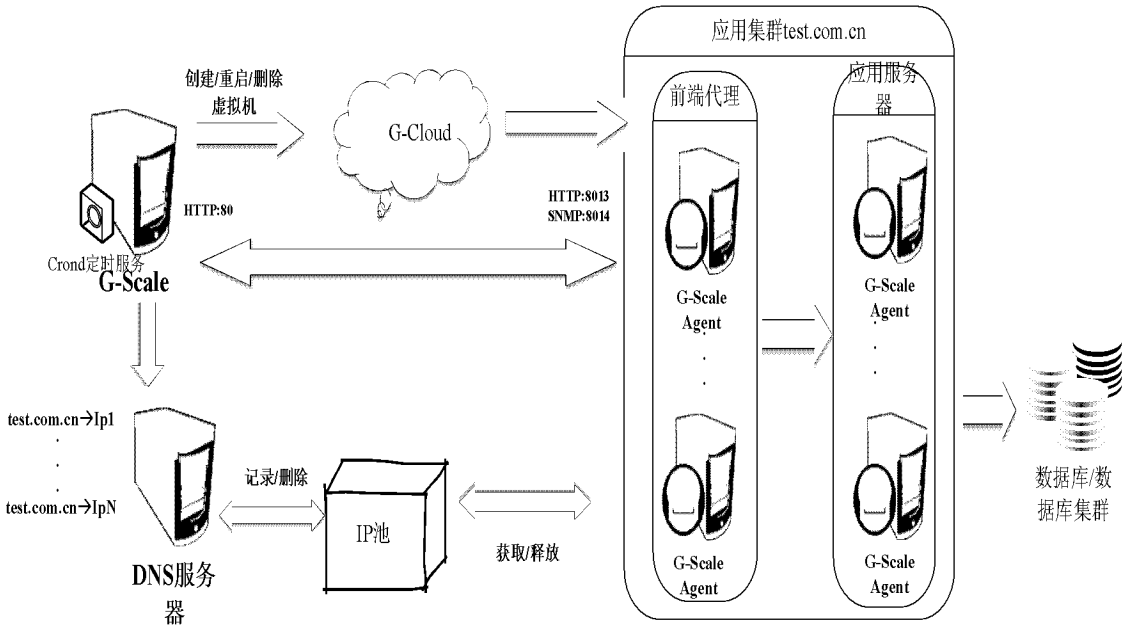

应用集群test.com.cn
前端代理
应用服务器
创建/重启/删除
虚拟机
G-Cloud
HTTP:80
HTTP:8013
SNMP:8014
Crond定时服务
G-Scale
G-Scale
Agent
G-Scale
Agent
G-Scale
Agent
G-Scale
Agent
test.com.cn→Ip1
test.com.cn→IpN
记录/删除
IP池
获取/释放
DNS服务器
数据库/数据库集群

图 1

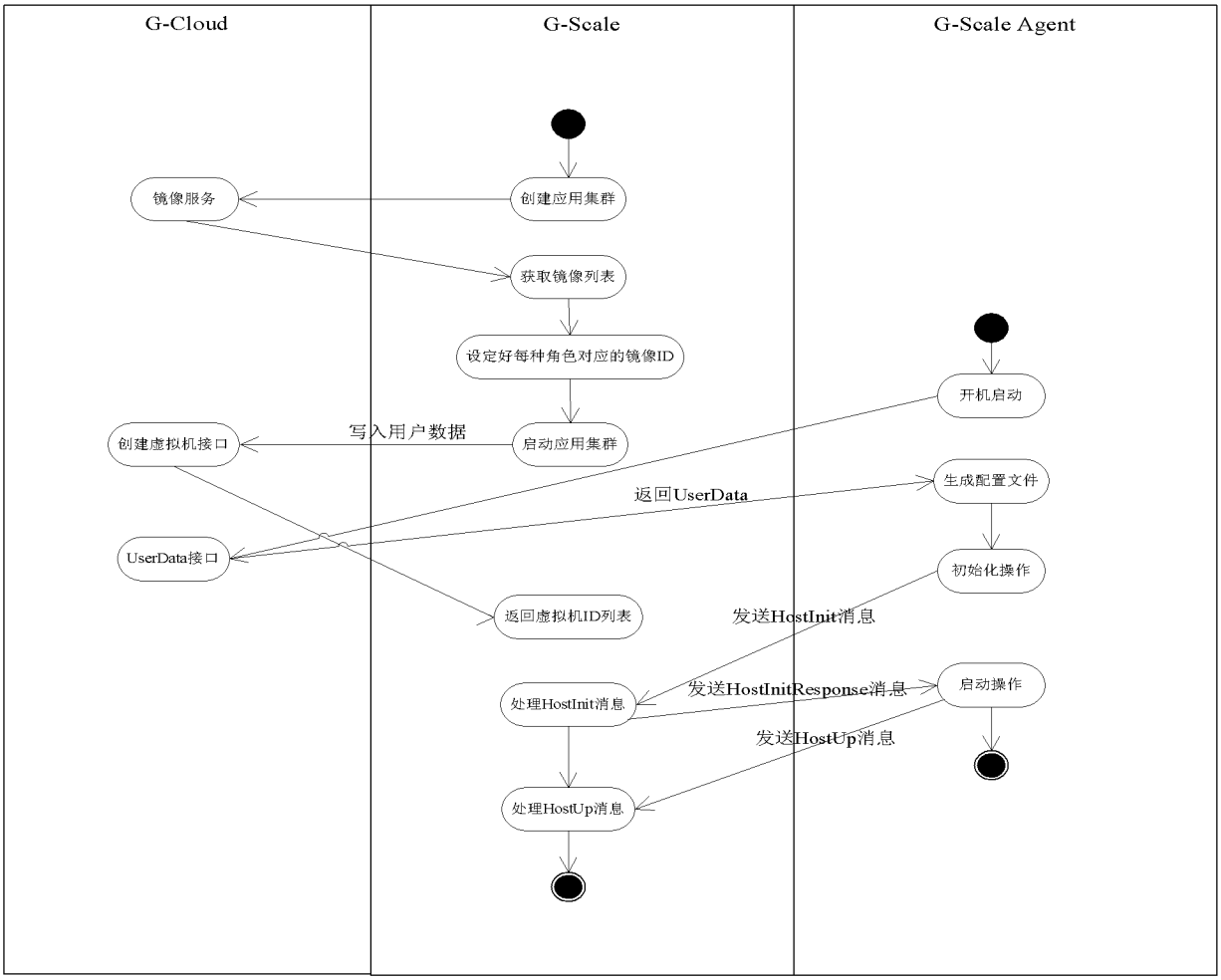

G-Cloud
G-Scale
G-Scale Agent
镜像服务
创建应用集群
获取镜像列表
设定好每种角色对应的镜像ID
开机启动
创建虚拟机接口
写入用户数据
启动应用集群
生成配置文件
返回UserData
UserData接口
初始化操作
返回虚拟机ID列表
发送HostInit消息
处理HostInit消息
发送HostInitResponse消息
启动操作
发送HostUp消息
处理HostUp消息

图 2