

「虚拟机」

基于虚拟机智能动态资源调度策略技术的研究

文/国网新疆电力公司信息通信公司 常春雷 杨大伟 陈 斌 向锋利

现有软硬件资源池中，各类业务系统分配的资源（CPU、内存和硬盘等）均通过人工分配，随意性较大，往往许多真实资源需求估算错误，造成资源服务能力不足，需要一种通过实际数据综合分析得出的各类软硬件资源的分配算法来提升各类资源的分配的科学性。重点研究了虚拟机智能动态资源调度策略技术，对软硬件资源池智能分析调控技术的研究对于提高资源利用率、节约能源和降低运营成本具有较为重要的意义。

目前，在服务器虚拟化技术的应用领域，多用静态的资源分配方法，虚拟机对负载变化不具有良好的适应性。在虚拟机集群的调度技术方面，大部分应用仍然使用传统物理集群下使用基于任务调度的策略。虚拟机动态迁移技术更是虚拟机化技术中的热点。因为虚拟机的动态迁移粒度大，很多采用虚拟机实现资源虚拟化的计算中心均利用虚拟机动态迁移技术实现负载均衡。由于任务型的应用不能覆盖所有类型的应用，基于任务调度的集群调度方法具有一定的局限性，并且调度粒度较大，难以发挥出虚拟机集群本身具有的特殊优势。在单个物理节点的虚拟机资源分配方面，大部分研究集中在虚拟机管理器VMM对虚拟机CPU时间片的分配算法、内存分页的管理和分配等方面，仅有少量文献提出了基于虚拟机负载的动态资源分配方法。单节点的动态资源分配具有速度快、开销小的特点，但是调度范围受限于单个节点，并且只能对虚拟机的CPU和内存资源进行动态资源分配。在多节点资源调度方面，VMware的商业产品vSphere DRS使用基于虚拟机迁移的手段进行多节点的负载平衡，它能为虚拟机提供透明的负载平衡，不对虚拟机中的应用进行任何更改，因而具有广泛的适用性。本文重点研究了虚拟机智能动态资源调度策略技术，对软硬件资源池智能分析调控技术的研究对于提

高资源利用率、节约能源和降低运营成本具有较为重要的意义。

自主软硬件资源池智能调控技术原理

不同的应用对不同资源的消耗和同一资源不同时间段的消耗是不同。而基于IaaS平台的应用部署是以专属虚拟机的形式部署，当虚拟机的资源不能满足应用的需求时，平台就要实时地增加分配给应用的资源，主要可以通过增加虚拟机的个数和对单台虚拟机进行资源的动态伸缩。对虚拟机资源的动态扩容，主要针对CPU，内存和磁盘。CPU动态调整则是对虚拟机占用宿主机CPU配额进行，其范围在预先设定的CPU配额上限之下波动，CPU配额调度基于全公平调度器（CFS）完全公平调度器。

CFS的设计思想是：在一个真实的硬件上模型化一个理想的、精确的多任务CPU，这个理想CPU模型运行在100%的负荷、在精确平等速度下并行运行每个任务，每个任务运行在 $1/n$ 速度下，即理想CPU任务运行数为 n ，每个任务的速度为CPU整个负荷的 $1/n$ ，由于真实硬件上，每次只能运行一个任务，这就得引入“虚拟运行时间”（virtual runtime）的概念，虚拟运行时间为一个任务在理想CPU模型上执行的下一个时间片（time slice）。实际上，一个任务的虚拟运行时间为考虑到运行

任务总数的实际运行时间。CFS背后的主要想法是维护为任务提供处理器时间方面的平衡。

内存则通过balloon机制来实现,通常来说,要改变客户机占用的宿主机内存,是要先关闭客户机,修改启动时的内存配置,然后重启客户机才能实现;而内存的气球技术可以在客户机运行时动态地调整它所占用的宿主机内存资源,而不需要关闭客户机。气球技术形象地在客户机占用的内存中引入气球的概念,气球中的内存是可以供宿主机使用的(但不能被客户机访问或使用),所以,当宿主机内存使用紧张,空余内存不多时,可以请求客户机回收利用已分配给客户机的部分内存,客户机就会释放其空闲的内存,此时若客户机空闲内存不足,可能还会回收部分使用中的内存,可能会换出部分内存到客户机的交换分区中,从而使得内存气球充气膨胀,从而让宿主机回收气球中的内存可用于其他进程(或其他客户机)。反之,当客户机中内存不足时,也可以让其他客户机的内存气球压缩,释放出内存气球中的部分内存,让客户机使用更多的内存。

虚拟机智能动态资源调度策略的实现

在传统集中式资源调度框架的基础之上提出了一个对等式的资源调度框架,并将基于现有的虚拟机动态迁移技术和负载均衡策略,提出一个新的虚拟机动态迁移方案,满足负载均衡和安全自愈的目的,总体设计流程如图1所示。

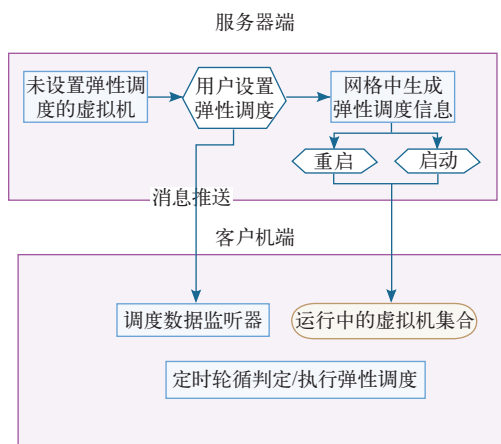


图1 总体设计流程示意图

1. 触发策略

所谓触发策略就是根据一定原则判断当前服务节点是否需要执行虚拟机的动态迁移。触发迁移策略主要由服务节点上的触发迁移模块来完

成。因为系统负载会偶尔的出现短时的上升,干扰触发时机的选择,所以为了确保一个短时间的负载上升不会触发动态迁移,需要设置一个合理的触发策略,设计一个基于二次指数平滑预测算法的触发迁移策略。

为了衡量节点的负载情况,我们需要将负载监听模块发送过来的资源使用率进行转化,得出具体的负载值。因为节点的负载指标较多,业务复杂,所以这里一律使用百分比来反映节点的负载情况。如20%负载、60%负载和100%负载等,这样单位同一直观,而不需要去考虑具体单位和具体的数值之间的关系。以内存使用率为例,定义当内存使用率为80%时,负载为100%,那么将80%定义为内存使用率的基准值(Mem_{base}),这样当内存使用率为50%时,计算出来的负载为62.5%,当内存使用率为100%的时候,计算出来的负载为125%。同样我们需要为CPU指标定义一些基准值作为负载的100%的值,在此我们也将该基准值定位80%(CPU_{base})。这样,就可以结合权重得到节点的负载值,为

$$L_{host} = \frac{\alpha L_{cpu}}{CPU_{base}} + \frac{(1-\alpha)L_{mem}}{Mem_{base}} \quad (1)$$

式中, L_{host} 即为当前节点的负载值; L_{cpu} 为节点的CPU使用率; L_{mem} 为节点的内存使用率; α 为CPU使用率在主机负载值中的权重,对应的 $1-\alpha$ 即为内存使用率在主机负载值中的权重。基于以上的介绍可以得出,服务节点的最大负载 L_{top} 为1.250。

2. 基于负载预测的触发流程

将触发迁移的阈值设为 M_h ,模块运行时开始观察每个周期的负载值,当有一个周期时刻 t 的负载值大于 M_h ,这个时候就开始连续观察该时刻后的 n 个周期的负载值,比如 $(L_{t+1}, L_{t+2}, L_{t+3}, \dots, L_{t+n})$,如果其中有大于或等于 J 个周期的负载值均大于 M_h ,则根据二次指数平滑法预测出的第 $n+1$ 个周期的负载值来判断是否触发迁移。

n 和 J 的值的选择会直接关系到虚拟机动态迁移的频率。如果 n 的值是固定的,则 J 值的下降可能会导致迁移频率的不断增加,导致迁移更加频繁,反之如果 J 值的增加会导致迁移的次数不断降低,使系统更加的保守。如果将 n 和 J 都设为1,就是简单的按照瞬时负载来决定虚拟机迁移条件,只要节点负载超过或低于设定阈值就开始虚拟机迁移。

二指平滑法(Second exponential smoothing

method) 不仅可以对时间序列值进行预测, 还能掌握时间序列的走势。因为系统负载是随着用户的使用量而波动的, 所以负载值会随机的上升或下降, 为了更好的判断当前系统的负载情况, 本文决定采用二次指数平滑法对系统负载以及负载的走势进行预测并通过预测值来决定是否触发虚拟机的动态迁移。

二次指数平滑法的预测模型为

$$L_{t+n} = a_t + nb_t \quad (2)$$

式中, S_{t+n} 为 $t+n$ 时刻的预测值; n 为预测值所在周期与当前周期的时间间隔; a_t 和 b_t 为预测参数。

$$a_t = 2S'_t - S''_t \quad (3)$$

$$b_t = \frac{\alpha}{1-\alpha}(S'_t - S''_t) \quad (4)$$

式中, S'_t 和 S''_t 分别为一次指数平滑值和二次指数平滑值。

$$S'_t = \alpha L_t + (1-\alpha)S'_{t-1} \quad (5)$$

$$S''_t = \alpha S'_t + (1-\alpha)S''_{t-1} \quad (6)$$

在使用二次指数平滑法来做预测的时候, 平滑参数 α 值的确定起着重要的作用。它的大小表示了当前负载值和预测负载值之间的影响关系。一般而言, 会采用最小均方差的原则来选取 α 的值, 因为 α 的取值范围为 0 到 1, 所以要在 0 到 1 之间取不同的 α 值进行观察, 然后分别计算出均方差, 取其中均方差最小的情况对应的 α 值作为实际使用时候的平滑参数。当然也可以通过长期的观察经验来取 α 的值, 本文便是通过对服务节点负载的持续观察, 来决定 α 的取值的。 α 不同的取值对平滑值的衰减速率有着不同的影响, α 越大, 平滑值衰减速率越快, 表示负载的预测值对之前负载值的依赖比较小, 说明本轮预测的负载值相对上一周期的负载值变化不大; 反之, α 的值越小, 平滑值衰减速率越快, 表示负载的预测值对以前负载值的依赖越大, 说明本轮预测的负载值相对上一周期的负载值变化较大。通过观察本文得到, 当 α 的值为 0.6 时, 预测的负载值相对与实际值的抖动越小, 预测值越准确。

3. 虚拟机策略设置

当确定需要执行虚拟机迁移后, 需要根据合理的虚拟机选择策略从源服务节点选择合适的待迁移虚拟机, 虚拟机选择策略由源服务节点的虚拟机选择模块完成, 虚拟机选择的原则是在尽量降低迁移开销的前提下, 释放更多的系统资源。弹性调控逻辑架构如图 2 所示。其中, 在服务器端

设置弹性调度, 配置弹性调度参数和调度结果展示, 并通过推送设置变化和客户端进行联系。在客户端端, 实现参数监听、虚拟机弹性调度数据缓存、虚拟机负载监控和弹性调度任务等功能。在弹性调控任务初始化阶段: 读取系统弹性调控参数, 初始化到缓存; 读取当前物理机上所有运行虚拟机的弹性调控数据, 初始化到弹性调控数据缓存; 设置所有虚拟机 CPU 初始调控配比。

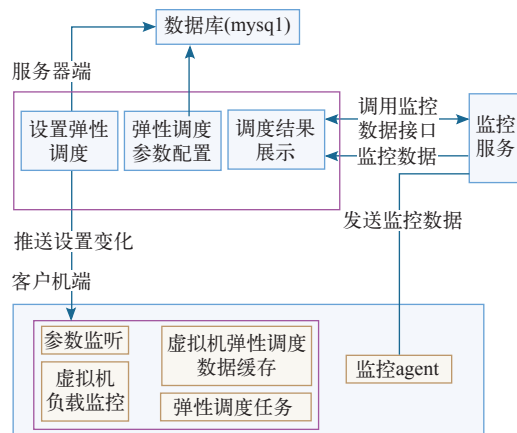


图2 弹性调控逻辑架构图

在弹性调控参数监听器设置阶段, 将虚拟机弹性调控参数更新到缓存。

在轮循执行调控任务阶段, 首先从缓存取得所有虚拟机弹性调控数据: 读取当前物理机上所有运行虚拟机, 与缓存中虚拟机比对: (有新增的虚拟机, 则从网络取得弹性调度数据, 放入缓存; 设置虚拟机 CPU 初始调控配比; 减小的虚拟机直接从缓存去除; 遍历缓存上虚拟机列表: 判定 CPU 调度数据存在, 执行“CPU 调度计算”, 执行“内存调度计算”。

弹性调控全局参数的设置和虚拟机弹性调控设置分别如图 3 和图 4 所示。

主要介绍了虚拟机动态迁移时的相关策略。触发策略通过平滑预测法判断是否触发虚拟机迁移, 避免了瞬时的负载上升或下降造成不必要的

```

CPU扩充条件: (vm.vcpu.elasticRate-cpu利用率)/vm.vcpu.elasticRate利用率<=x%
(x取值范围:5<=x<=15, 正整数)x默认=10
CPU缩减条件: (vm.vcpu.elasticRate-cpu利用率)/vm.vcpu.elasticRate>=y%
(y取值范围:35<=y<=45, 正整数)y默认=40
扩充量: vm.vcpu.elasticRate*n%, (n取值范围:2<=n<=10)n默认=10

内存扩充条件: mem>=x% (x取值范围:65<=x<=90, 正整数, 默认=80)
内存缩减条件: mem<=m% (m取值范围:10<=m<=45, 正整数)默认=35
内存单次扩充/缩减量 = 当前量*y%(取整)
(y取值范围: 2<=y<=10, 正整数), y默认=10

CPU最小缩减百分比: p% (p取值范围:10<=p<=50, 正整数)默认=20
内存最小缩减百分比: S% (s取值范围: 20<=s<=50, 正整数)默认=30
弹性调度频率 x 秒, 正整数
    
```

图3 弹性调控全局参数设置图

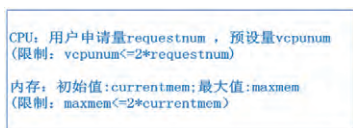


图4 虚拟机弹性调控设置图

虚拟机迁移;虚拟机选择策略为了达到降低源服务节点负载的同时减少动态迁移的开销,充分考虑了宿主机的资源使用情况。

4. 智能化资源调度主要模块的实现

上文分别介绍了实现资源调度策略的资源调

度框架和虚拟机动态迁移策略。下文将基于上文的内容介绍智能化资源调度主要模块的实现。

本地迁移引擎的实现。本地迁移引擎包括节点负载和虚拟机负载监听模块、触发迁移模块、虚拟机选择模块、虚拟机调度模块和网络通信模块,以下小节主要对前面四个模块的实现做详细介绍,网络通信模块的实现将放在与中心控制引擎的网络通信模块一起介绍。本地迁移引擎实现的主要类如图5所示。

网络通信模块的实现。通过前面的介绍我们

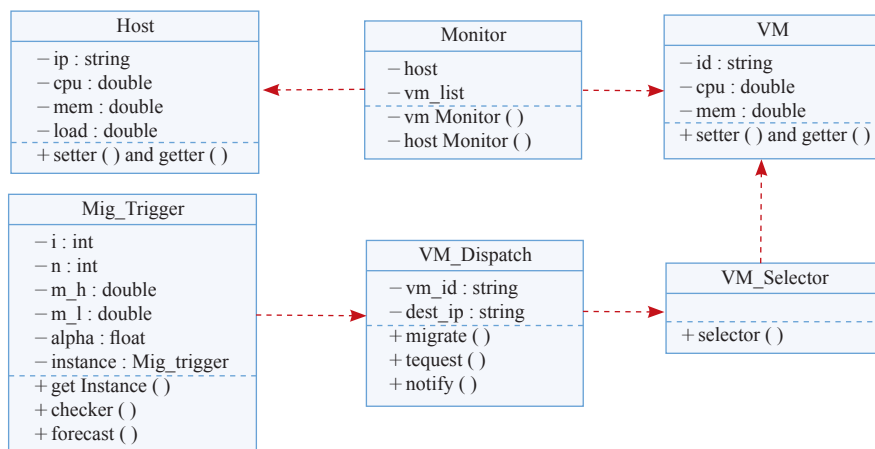


图5 本地迁移引擎类图

知道,本地迁移引擎和中心控制引擎之间要不时的进行网络通信,为了简化本地迁移引擎和中心控制引擎其它模块的操作流程,让它们能够把主要工作放在业务逻辑的处理,本文在这里将网络通信独立出来作为一个模块。本论文的网络通信模块基于MINA框架实现。MINA是一个基于Java NIO (Non-Blocking IO) 的网络框架,该框架能够同时在客户端和服务端发挥很好的通信效果。因为MINA对底层的网络通信进行了很好的封装,所以当我们使用MINA来实现应用的时候,不用去考虑底层诸多复杂的参数和不同的应用场景。

接收功能模块的实现。本地迁移引擎需要向中心控制引擎发送负载信息和请求信息,所以需要网络通信模块提供主动消息发送功能;因为中心控制引擎需要定时发送心跳消息,所以中心控制引擎的网络通信模块也要体现消息发送功能。同理,本地迁移引擎和中心控制引擎的网络通信模块也都需要实现消息接收功能。接收功能类似于网络通信的服务器端,需要不断的在特定的端口上进行监听,然后处理接收到的数据。

介绍了本文虚拟机调度策略主要模块的实

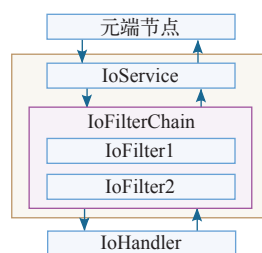


图6 网络通信模块实现框架层次图

现,包括本地迁移引擎的实现、网络通信模块等的实现。

结束语

虚拟机动态迁移技术更是虚拟机化技术中的热点。因为虚拟机的动态迁移粒度大,很多采用虚拟机实现资源虚拟化的计算中心均利用虚拟机动态迁移技术实现负载均衡。重点研究了虚拟机智能动态资源调度策略技术,对软硬件资源池智能分析调控技术的研究对于提高资源利用率、节约能源和降低运营成本具有较为重要的意义。EA

(收稿日期:2016.09.26)

(编辑:白文亭)