

基于虚拟机迁移的虚拟机集群负载均衡策略研究

张 创, 谷建华

(西北工业大学 计算机学院, 陕西 西安 710072)

摘 要: 针对虚拟机集群资源负载不均衡的问题, 提出了一种多策略与预测机制相结合的、基于虚拟机迁移的负载均衡方案. 对导致系统过载的不同类型的资源采用不同的均衡策略, 能够更加有效地使系统恢复正常. 采用预测机制, 消除瞬时峰值导致的不必要的迁移, 可以减少传统的基于单阈值策略中虚拟机的迁移频率. 在触发迁移后采用概率转发方式选择目的节点, 能够有效解决群聚冲突问题. 实验结果表明, 本方案能够有效地实现系统的负载均衡.

关键词: 负载均衡; 虚拟化; 迁移; 预测; 资源调度

中图分类号: TP315

文献标识码: A

文章编号: 1000-7180(2014)04-0079-04

A Load Balancing Strategy in Virtual Machine Cluster Based on Live Migration of Virtual Machine

ZHANG Chuang, GU Jian-hua

(School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: To solve the problem of unbalancing resource load status in virtual machine cluster, this paper describes a load balancing strategy which based on live migration and combines multi-strategy mechanism with the prediction mechanism. It adopts different balancing strategys for different types of resources to make the system return to normal effectively, and uses the prediction mechanism to eliminate unnecessary migration caused by instantaneous peak, so it can reduce the frequency of virtual machine migration in traditional single threshold strategy. After triggering migration, it uses probability-weighted approach to select the target node, which resolves the problem of accumulation conflicts. Experimental results show that the method can achieve load balancing.

Key words: load balancing; virtualization; live-migration; prediction; resource scheduling

1 引言

近年来, 虚拟化技术在云计算领域备受关注^[1]. 虚拟机不仅提供一个具有逻辑独立的, 安全和高效的计算资源, 实现最大化的利用硬件资源, 而且可以在多个物理主机之间动态迁移, 有利于系统负载均衡的实现, 而负载均衡是保证虚拟机集群正常工作的基石.

任务的不确定性和宿主机处理能力的差异性都会导致系统中宿主机的负载失衡. 现阶段, 大多数负载均衡策略的实现都是基于阈值的, 当节点的负载超过上限阈值时, 触发迁移, 将部分虚拟机迁移到轻

载节点上, 实现负载均衡, 如文献[2-4]. 然而这种情况下, 一个很小的瞬时峰值都会触发一次不必要迁移, 造成网络拥塞, 严重影响系统的性能. 文献[5]采用预测机制, 当节点的负载超过上限阈值时, 根据节点负载的历史数据判断下一时刻节点的负载情况, 决定是否迁移. 此方案虽然可以避免瞬时峰值触发的迁移, 但是在节点处于过载时才进行预测, 使得节点处于过载的时间更长, 性能损失较大, 同时它将各资源按权相加, 以其和值判断节点是否过载, 如果各权值选择不当, 可能会导致和值小于阈值, 而其中某一种资源负载过重, 也会影响系统性能.

本文提出一种多策略与预测机制相结合的方

案,将宿主机中某类资源的负载分成不同区域,对不同区域采用不同的策略,预测算法判断资源的负载趋势.同时,采用概率转发方式选择迁移的目标节点,解决多节点基于相同的负载信息可能选择同一个节点机作为目标节点,导致被选中的目标节点负载急剧增加,造成群聚冲突问题.

2 方案分析与设计

本文的负载均衡方案的实现分三个阶段,即判断迁移,选择欲迁移的虚拟机和选择目的主机.

2.1 判断迁移

在此阶段中,首先对系统各资源的不同使用情况进行分区,在不同区中执行不同的方案,然后对资源使用率进行预测,判断是否进行迁移.

2.1.1 资源负载分区

我们将资源的不同负载划分为如图 1 所示的四个分区,其中的变量含义为: S_{cpu} 表示宿主机 CPU 使用率, S_{mem} 表示宿主机内存使用率, S_{net} 表示宿主机网络带宽使用率, Q_1 表示轻载的阈值, Q_2 表示预警阈值, Q_3 表示过载阈值.

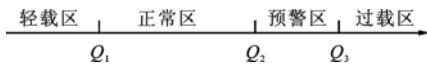


图 1 资源负载分区

轻载区:满足 $(S_{cpu} < Q_1) \& (S_{mem} < Q_1) \& (S_{net} < Q_1)$, 此时可以认为宿主机的整体负载很轻,故可以迁移出此物理机上的所有虚拟机,并关闭此物理机,以节约能源,同时将其加入到系统中已关机物理机队列 Q_{down} .

正常区:满足 $(Q_1 \leq S_{cpu} < Q_2) \& (Q_1 \leq S_{mem} < Q_2) \& (Q_1 \leq S_{net} < Q_2)$, 此时宿主机整体负载处于正常状态,不需要对此物理机进行操作.

预警区:满足 $(Q_2 \leq S_{cpu} < Q_3) \vee (Q_2 \leq S_{mem} < Q_3) \vee (Q_2 \leq S_{net} < Q_3)$ 其中一个条件,则触发负载均衡机制,预测负载的趋势,判断是否迁移.

过载区:满足 $(S_{cpu} \geq Q_3) \vee (S_{mem} \geq Q_3) \vee (S_{net} \geq Q_3)$ 其中一个条件,则物理机处于过载状态,同时进行虚拟机的迁移.

2.1.2 预测算法

本文采用时间序列预测技术中的自回归模型 $AR(n)$ 预测将来的负载值^[6].该模型使用 n 个过去的、按时间序列排列的观测值对未来做出预测.例如观测的负载值序列: $y_1, y_2, \dots, y_t$. 对于这个给定的时间序列,利用 p 阶自回归模型 $AR(p)$ 预测第 t 个时刻的负载值,则第 t 个时刻的预测值 $\hat{y}_t$ 如下:

以 CPU 使用率为例,每隔 T 时间观察 CPU 使

用率,使用过去的观察值预测将来值.

$$\hat{y}_t = \phi_1 y_1 + \phi_2 y_2 + \dots + \phi_p y_{t-p} + \epsilon \quad (1)$$

式中, ϵ 为误差; $\phi_1, \phi_2, \dots, \phi_p$ 为时间序列的参数.通过公式(1)预测未来 n 个负载值,如果至少有 k 个值大于 Q_3 ,则触发迁移,否则继续预测.

2.2 选择欲迁移的虚拟机

对于由不同类型资源触发的迁移,选择下面不同的策略.

(1) 由 CPU 触发的迁移:此时需综合考虑虚拟机的 CPU 的使用率和内存占用情况.虚拟机对内存的占用量越大,迁移的代价也越大,但同时,迁移也释放原节点更多的内存空间.鉴于此,当触发迁移时,如果宿主机内存使用率也过高,比如 $S_{mem} \geq Y_{mem}$,则选择迁移内存使用率 $V_m(i)$ 和 CPU 使用率 $V_c(i)$ 都较高的虚拟机;反之,则迁移内存使用率低而 CPU 使用率高的虚拟机.此处, Y_{mem} 是自定义的物理机内存使用率,用以判断选择虚拟机的方法, $V_m(i)$ 表示虚拟机 i 占用物理机内存的比率, $V_c(i)$ 表示虚拟机 i 占物理机 CPU 的比率.

(2) 由内存触发的迁移:如果此时 $S_{cpu} \geq Y_{cpu}$,则选择内存使用率和 CPU 使用率都较高的虚拟机;反之,选择内存使用率高, CPU 使用率低的虚拟机.

(3) 由网络带宽触发的迁移:此时只选择带宽消耗最大的虚拟机进行迁移.

2.3 选择目的节点

根据候选节点的可用资源量,计算其被选中的概率,可用资源越多,被选中的概率越大.

2.3.1 相关变量的定义:

CPU 相关变量定义, f 表示物理机 CPU 的主频; N_{cpu} 表示 CPU 核心数; $R_{pm_cpu_max}$ 表示物理机 CPU 的最大能力; $R_{pm_cpu_used}$ 表示物理机已使用的 CPU 能力; $R_{vm_cpu}(i)$ 表示虚拟机 i 的 cpu 能力,各变量计算公式如下:

$$R_{pm_cpu_max} = f \times N_{cpu} \quad (2)$$

$$R_{pm_cpu_used} = R_{pm_cpu_max} \times S_{cpu} \quad (3)$$

$$R_{vm_cpu}(i) = f \times N_{cpu} \times V_c(i) \quad (4)$$

内存相关变量定义, $R_{pm_mem_max}$ 表示物理机内存最大值; $R_{pm_mem_used}$ 表示物理机已使用内存;

$R_{vm_mem}(i)$ 表示虚拟机 i 的实际物理内存使用量,计算公式(5)如下:

$$R_{vm_mem}(i) = R_{pm_mem_max} \times V_m(i) \quad (5)$$

网络带宽相关变量定义, $R_{pm_net_max}$ 表示物理机的最大网络带宽; $R_{pm_net_used}$ 表示物理机当前已使用网络带宽; $R_{vm_net}(i)$ 表示虚拟机 i 的网络带宽.

使用向量 $\mathbf{V}$ 、 $\mathbf{W}$ 分别表示欲迁移虚拟机的不同

资源需求向量和目的物理机可以分配的不同资源向量. 向量定义如下:

$$\mathbf{V} = \langle R_{vm_cpu}(i), R_{vm_mem}(i), R_{vm_net}(i) \rangle$$

$$\mathbf{W} = \langle R_{pm_cpu_r}, R_{pm_mem_r}, R_{pm_net_r} \rangle$$

其中向量 $\mathbf{W}$ 中各变量计算公式如下:

$$R_{pm_cpu_r} = R_{pm_cpu_max} \times Q_2 - R_{pm_cpu_used} \quad (6)$$

$$R_{pm_mem_r} = R_{pm_mem_max} \times Q_2 - R_{pm_mem_used} \quad (7)$$

$$R_{pm_net_r} = R_{pm_net_max} \times Q_2 - R_{pm_net_used} \quad (8)$$

2.3.2 计算候选节点被选中的概率

定义变量 S_i :

$$S_i = k_1 \times \frac{R_{pm_cpu_r}(i)}{\sum_{i=1}^n R_{pm_cpu_r}(i)} + k_2 \times \frac{R_{pm_mem_r}(i)}{\sum_{i=1}^n R_{pm_mem_r}(i)} + k_3 \times \frac{R_{pm_net_r}(i)}{\sum_{i=1}^n R_{pm_net_r}(i)} \quad (9)$$

式中, $k_1 + k_2 + k_3 = 1$, k_1, k_2, k_3 表示各资源的权重.

则节点被选中的概率 $P_i = \frac{S_i}{\sum_{i=1}^n S_i}$.

例如节点集合 $N = \{N_1, N_2, N_3, N_4\}$, 节点被选中的概率 $P = \{40\%, 30\%, 20\%, 10\%\}$, 用一个随机函数生成一个 $[0, 1]$ 之间的数, 然后根据随机数落在哪个目标节点的概率空间中, 最后确定迁移到哪个目标节点, 如图 2 所示.

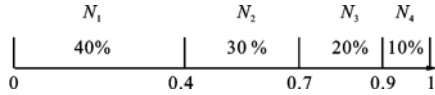


图2 节点的概率空间图

如果正在运行的物理节点中没有满足资源要求的物理节点, 则系统从已关机物理机队列 Q_{down} 中选择一个物理机启动.

3 实验结果

为了验证本文方法的性能, 进行了3组实验. 测试环境包含5台宿主机, 各宿主机包含不同数量及不同配置的虚拟机. 本实验中设置 Q_2 阈值 0.85, Q_3 阈值为 0.95, 建立的预测模型为 AR(6).

3.1 CPU 负载测试

我们在不同物理机的虚拟机上运行不同任务量的CPU型作业, 然后记录某时刻迁移发生前后各物理机的CPU负载情况, 实验结果如图3所示.

从图3可以看出, 在迁移发生前, 主机2处于过载区, 主机3、5处于预警区. 其中, 主机2、3发生迁移, 因为主机2的负载大于 Q_3 从而触发迁移, 预测算法判断出主机3的负载将会大于 Q_3 , 从而也触发迁移; 而通过预测算法得出主机5的负载呈下降趋势, 不触发迁移, 从迁移后的曲线可知, 主机5的负

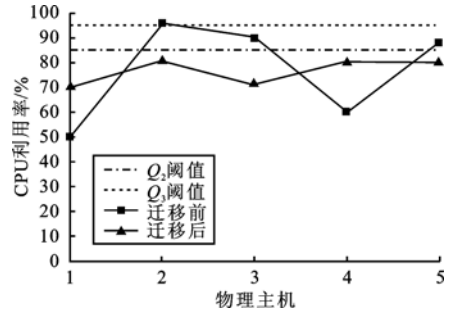


图3 迁移前后物理主机CPU利用率
载确实降低了, 因此减少了一次不必要的迁移.

3.2 内存负载测试

我们在不同虚拟机上运行不同任务量的内存型作业, 然后记录某时刻迁移发生前后各物理机的内存负载情况, 实验结果如图4所示.

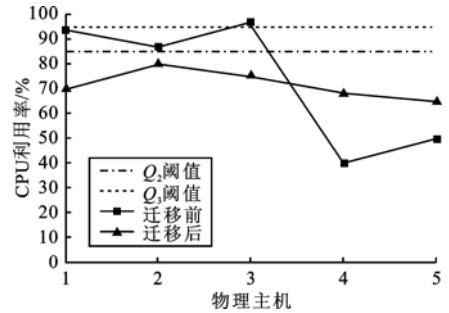


图4 迁移前后物理主机内存利用率

从图4可以看出, 迁移发生前主机3处于过载区, 主机1、2处于预警区. 其中, 主机1、3发生迁移, 预测算法判断出主机1的负载将会超过 Q_3 , 故触发迁移, 主机3的负载大于 Q_3 , 触发迁移; 根据预测算法得出主机2的负载有下降趋势, 故不迁移, 由迁移后的曲线可知, 主机2的负载确实降低了, 避免了不必要的迁移.

3.3 带宽负载测试

我们在不同虚拟机上运行不同任务量的带宽型作业, 然后记录某时刻迁移发生前后各物理机的带宽负载情况, 实验结果如图5所示.

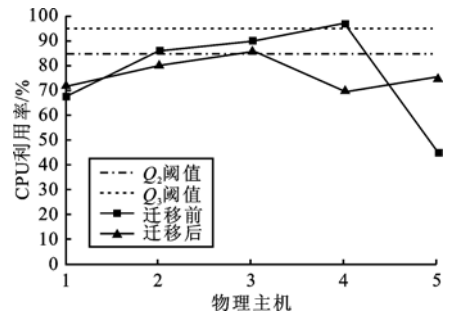


图5 迁移前后物理主机带宽负载

从图5可以看出, 迁移发生前, 主机4处于过载

区,主机 2、3 处于预警区.其中,主机 4 发生迁移;根据预测算法得出主机 2、3 的带宽负载呈下降趋势,不会超过 Q3,所以不发生迁移.从迁移前后曲线对比可以看出,主机 2、3 的带宽负载确实下降了.

4 结束语

本文提出了一种采用多策略与预测机制相结合的负载均衡方案,对于导致系统过载的不同类型的资源采用不同的均衡策略,由某种资源触发的迁移,在欲迁移虚拟机的选择方面,不仅考虑了本资源的负载,也考虑了其他资源的负载,使得本次迁移不仅解决了本资源的过载,也使得其他资源过载的可能性降低.同时,预测机制的引入,消除了瞬时峰值导致的不必要的迁移,有效的减少了基于单阈值策略中虚拟机的迁移频率,减少了系统开销.在目的主机的选择的方面,采用概率转发方式,避免了多虚拟机同时迁移到同一主机所引起的聚群效应.本文的方案可以有效的实现系统的负载均衡.

参考文献:

- [1] Paul Barham, Boris Dragovic, Keir Fraser, et al. Xen and the art of virtualization[C]//Proceedings of the nineteenth ACM symposium on Operating Systems Principles (SOSP19). New York: ACM Press, 2003:164-177.
- [2] Khanna G, Beaty K, Kar G, et al. Application performance management in virtualized server environments[C]// Proc. of Network Operations and Management Symp. Vancouver: IEEE, 2006.
- [3] Timothy Wood, Prashant Shenoy. Black-box and gray-box strategies for virtual machine migration[C]//Proceedings of the 4th USENIX Symposium Conference on Networked Systems Design & Implementation. San Diego: IEEE Press, 2007.
- [4] VM. Vmware DRS[EB/OL]. [2013-03-02]. <http://www.vmware.com/cn/products/datacenter-virtualization/vsphere/drs-dpm.html>, 2012.
- [5] Weiming Zhao, Zhenlin Wang. Dynamic memory balancing for virtual machines[J]. Operating Systems Review, 2009, 43: 37-47.
- [6] Box G P, Jenkins G M, Reinsel G C. Time Series Analysis: Forecasting and Control[M]. 北京:人民邮电出版社, 2005.
- [7] Liao Yuan-Hsin, Li Gwo-Long, Tian-Sheuan Chang. 135-MHz 258-K Gates VLSI Design for All-Intra H. 264/AVC Scalable Video Encoder [J]. IEEE TRANS. on CSVT, 2012, 22(2):272-281.
- [8] Wang X, Cui X X, Yu D S. A parallel intra prediction architecture for H. 264 video decoding [C]//2009 IEEE 8th International Conference on ASIC. Changsha, Hunan: IEEE Press, 2009.
- [9] 吴斌,郭树旭,王明江,等.一种可重构的 H. 264 帧内预测硬件结构[J]. 微计算机信息, 2009, 25(3):286-288.
- [10] Xu K, Choy C S. A power-efficient and self-adaptive prediction engine for H. 264/AVC decoding[J]. Very Large Scale Integration (VLSI) Systems, 2008, 16(3): 302-313.
- [11] Wu L G, Zhang D L, Du G M, et al. A 4x4 pipelined intra frame decoder for H. 264 [C]//Anti-counterfeiting, Security, and Identification in Communication. Hong Kong: IEEE Press, 2009.
- [12] He X, Zhou D J, Zhou J J, et al. A new architecture for high performance intra prediction in H. 264 decoder [C]//Intelligent Signal Processing and Communication Systems. Kanazawa: IEEE Press, 2009.
- [13] Lai Y F, Liu T M, Li Y, et al. Design of an intra predictor with data reuse for high-profile H. 264 applications [C]//Circuits and Systems. Taipei: IEEE Press, 2009.
- [14] He X, Zhou D J, Zhou J J, et al. High profile intra prediction architecture for H. 264 [C]//SoC Design Conference. Busan: IEEE Press, 2009.
- [15] Lin T A, Wang S Z, Liu T M, et al. An H. 264/AVC decoder with 4x4-block level pipeline[C]//Circuits and Systems. Kobe: IEEE Press, 2005.

作者简介:

张 创 男,(1988—),硕士.研究方向为云计算.

谷建华 男,(1965—),博士,博士生导师.研究方向为高性能计算和云计算.

(上接第 78 页)

- [6] Liao Yuan-Hsin, Li Gwo-Long, Tian-Sheuan Chang. 135-MHz 258-K Gates VLSI Design for All-Intra H. 264/AVC Scalable Video Encoder [J]. IEEE TRANS. on CSVT, 2012, 22(2):272-281.
- [7] Wang X, Cui X X, Yu D S. A parallel intra prediction architecture for H. 264 video decoding [C]//2009 IEEE 8th International Conference on ASIC. Changsha, Hunan: IEEE Press, 2009.
- [8] 吴斌,郭树旭,王明江,等.一种可重构的 H. 264 帧内预测硬件结构[J]. 微计算机信息, 2009, 25(3):286-288.
- [9] Xu K, Choy C S. A power-efficient and self-adaptive prediction engine for H. 264/AVC decoding[J]. Very Large Scale Integration (VLSI) Systems, 2008, 16(3): 302-313.
- [10] Wu L G, Zhang D L, Du G M, et al. A 4x4 pipelined intra frame decoder for H. 264 [C]//Anti-counterfeiting, Security, and Identification in Communication. Hong Kong: IEEE Press, 2009.
- [11] He X, Zhou D J, Zhou J J, et al. A new architecture for high performance intra prediction in H. 264 decoder [C]//Intelligent Signal Processing and Communication Systems. Kanazawa: IEEE Press, 2009.
- [12] Lai Y F, Liu T M, Li Y, et al. Design of an intra predictor with data reuse for high-profile H. 264 applications [C]//Circuits and Systems. Taipei: IEEE Press, 2009.
- [13] He X, Zhou D J, Zhou J J, et al. High profile intra prediction architecture for H. 264 [C]//SoC Design Conference. Busan: IEEE Press, 2009.
- [14] Lin T A, Wang S Z, Liu T M, et al. An H. 264/AVC decoder with 4x4-block level pipeline[C]//Circuits and Systems. Kobe: IEEE Press, 2005.

作者简介:

兰旭光 男,(1976—),博士,副教授.研究方向为图像视频内容理解、图像视频编解码及 VLSI 设计.

李兴玉 女,(1988—),硕士研究生.研究方向为视频编解码的 VLSI 设计.

温 灏 男,(1989—),硕士研究生.研究方向视频编解码的 VLSI 设计.

王志刚 男,(1988—),硕士研究生.研究方向为图像和视频内容分析.