

基于概率预测的虚拟机动态迁移机制改进

方义秋,陈玉洋,葛君伟

(重庆邮电大学 计算机科学与技术学院,重庆 400065)

摘 要: 提出在基于概率预测的 pre-copy 算法的基础上,引入带宽自适应算法,当脏页率突然增大时,带宽自适应算法能够自动提高迭代传输带宽,缩短传输时间. 实验结果表明,在脏页率较高和脏页率变化较大时,该策略能够明显缩短迭代传送时间和停机时间,提高迁移性能和业务服务质量.

关键词: 云计算;虚拟机;动态迁移;脏页率;带宽自适应

中图分类号: TP31

文献标识码: A

文章编号: 1000—7180(2016)11—0176—05

Improvement of Live Migration Mechanism for Virtual Machine Based on Pre-Copy

FANG Yi-qiu, CHEN Yu-yang, GE Jun-wei

(College of Computer, Chongqing University of Post and Telecommunication, Chongqing 400065, China)

Abstract: An Adaptive Bandwidth Algorithm is introduced in the pre-copy algorithm based on probability prediction. When dirty pages rate suddenly increases to a larger value, that algorithm could improve the transmission bandwidth to shorten transmission time. Experimental results show that the rate of dirty pages increase suddenly or is great, that strategy can improve the transmission bandwidth, shorten the total time of iterative transfer and downtime, so as to improve migration performance and quality service.

Key words: cloud computer; virtual machine (VM) live migration; dirty page rate; bandwidth adaptation

1 引言

云计算作为一种便捷的、可用的、按需的、网络可访问的资源,为广大政务、企业、个人提供服务. 不同的应用对资源的需求是不同的,云计算通过虚拟多个虚拟主机来装载不同的应用服务,并用虚拟化的技术来管理这些共享资源,并自适应地对资源进行分配和调度,以此管理各种物理资源,从而高效绿色地对云计算平台中的资源和服务进行管理^[1]. 虚拟机动态迁移 (Live Migration) 技术^[2],是在将虚拟机从源主机迁移到目的主机的过程中不影响虚拟机上的应用程序的情况下向外界提供服务,使之继续运行,从而达到在线系统维护、动态负载均衡等目的,实现高效、绿色的管理资源^[3]. 虚拟机动态迁移的总体性能取决于内存迁移性能的好坏,而内存迁移的主要性能指标为迭代拷贝

时间和宕机时间. 目前,广泛使用的内存迁移技术是预拷贝 (Pre-Copy) 机制^[2],以循环迭代的方式将内存页面从源虚拟机拷贝到目的虚拟机. 在较低负载时,Pre-Copy 算法表现较好,实现了透明的迁移,但是在负载较高时表现却不理想.

2 相关工作

文献^[2]提出的预拷贝 (Pre-Copy) 模型以迭代拷贝的方式对虚拟机内存进行动态迁移,第 N 轮传送第 $N-1$ 轮中产生的脏页,直到可写工作集小于预定的阈值,暂停虚拟机,传送可写工作集的所有页面. 预拷贝算法实现了虚拟机透明的动态迁移. 由于部分内存页面的频繁更改,产生脏页,导致传送了大量重复页面,影响性能.

针对 Pre-Copy 算法的缺点,文献^[4]对其进行

了优化,引入了马尔科夫模型来预测页面变脏的概率.该策略在一定程度上减少脏页面的重复传送,提高了性能;但当内存页面突然被密集读写时,概率预测可能阶段性失效,可能会造成迁移失败.

文献[5]提出,在内存迁移前,通过 Lempel-Ziv 算法来压缩待传输页面,并且在 Xen 虚拟机的迁移代码中使用贝叶斯条件概率来减少传输内存页的数量,在一定程度上减少了数据传输总量,但是并没有减少数据传输次数,数据传输前后都需要一定的时间来压缩和解压缩内存.因此,在总迁移时间上并没有明显缩短迁移总时间.

本文在基于马尔科夫概率预测的 Pre-Copy 算法的基础上,引入了带宽自适应算法,当内存页面突然被密集地读写时,脏页率大幅提高,预测的脏页概率与实际有较大的差距,为了缩短迁移时间,可以通过引入带宽自适应算法提高传送带宽,通过提高带宽来缩短迁移总时间和停机时间.

3 基于马尔科夫概率预测的带宽自适应算法

虚拟机动态迁移的性能指标包括迁移总时间和宕机时间.根据虚拟机迁移过程,可得迁移总时间和宕机时间计算公式如下:

$$T_{\text{total}} = T_{\text{init}} + T_{\text{res}} + \sum_{i=1}^{n-1} \text{Pre-copy}_i + T_{\text{down}} + T_{\text{active}} \quad (1)$$

$$T_{\text{down}} = T_{\text{stopcopy}} + T_{\text{commit}} \quad (2)$$

由式(1)和(2)可知,迁移总时间 T_{total} 由迁移初始化时间 T_{init} 、资源预定时间 T_{res} 、Pre-Copy 算法的 n 轮迭代时间、停机拷贝时间 T_{stopcopy} 、提交确认时间 T_{commit} 和 T_{active} 激活组成.停机总时间 T_{down} 由 T_{stopcopy} 和 T_{commit} 组成, T_{init} 、 T_{res} 、 T_{commit} 提交确认和 T_{active} 激活都是固定的开销.为了缩短迁移总时间和停机时间,必须从 Pre-Copy 算法和 T_{stopcopy} 入手.而基于马尔科夫概率预测的 Pre-Copy 算法就是从 Pre-Copy 算法的 n 轮迭代时间入手.

3.1 基于马尔科夫概率预测的迁移机制

针对传统迁移机制存在的缺点,文献[4]对其做了优化设计,引入了马尔科夫概率模型来对脏页出现的概率进行预测.该策略添加了概率预测模块,其功能为在迭代前预测内存页面被修改的概率,根据概率判断是否将其添加到工作集中.

在对 VM 进行迁移时,虚拟机监控器(Virtual Machine Monitor, VMM)中的影子页表能够捕捉到被读写的内存页,并通过脏页位图[8]来记录这些

变脏的页面,概率预测模块根据脏页位图中历史记录来预测内存页被修改的概率.设拷贝总轮数为 n ,每轮发送的内存脏页数 D_i ,发送每轮内存页面的带宽为 BW_i ,每轮迭代使用的时间为 T_i ,则有

$$T_i = \frac{D_i}{BW_i}, < i < n \quad (3)$$

在最后一轮,停止 VM,并进入停机拷贝阶段,由式(2) 停机拷贝时间 T_{down} ,则

$$T_{\text{down}} = \frac{D_i}{BW_i} + T_{\text{commit}}, i = n \quad (4)$$

综合式(1) ~ (4) 可得:

$$T_{\text{total}} = T_{\text{init}} + T_{\text{res}} + \sum_{i=1}^{n-1} T_i + \frac{M_i}{BW_i} + T_{\text{active}} \quad (5)$$

迁移执行模块接收到迁移信号后,便开始监听内存页的变化.概率预测模块会根据脏页位图来预测页面变脏的概率,被修改过的页面再次被修改的概率高,将其添加到可写工作集中,而预测的概率低时,便将这些页面传输到目的虚拟机.重复执行,直到符合停止迭代的条件[6],迁移执行模块便启动冻结模块和目标域唤醒模块.

基于马尔科夫概率预测的迁移策略,较大地缩短了传送数据总量和迁移时间.但是,当运行在虚拟机上的应用突然需要大量读写内存页面时,概率预测就会有较大的误差,降低了迁移性能.

3.2 基于马尔科夫概率预测的带宽自适应机制

上节提到基于马尔科夫概率预测算法的缺点,本文对其进行改进,如图 1 所示.改进后的机制主要增加了带宽自适应模块.

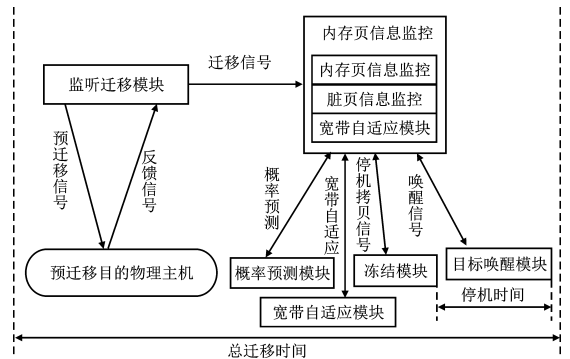


图 1 改进后的迁移机制

在实际应用中,不同的应用服务占用计算机的主要物理资源是不同的.因此,在迁移过程中,每一轮发送的脏页面数量是不断变化的.为此,引入了带宽调整系数 $a^{[7]}$, $a \in [0, 1]$,不仅考虑到了虚拟机中当前业务所使用的带宽量,而且也考虑到了迭代中反映业务特征的脏页率.通过式(6)计算第 k 轮迭代

拷贝过程中虚拟机中的业务使用的带宽量.

$$B_{bw} = (1 - a) \times \text{avg}(\sum_{i=1}^{k-1} U_{dirty_i}) + a \times U_{dirty_k} \quad (6)$$

U_{dirty} 为历史脏页率, 带宽调节系数 a 体现了在对带宽进行分配时, 历史数据和当前数据对系数 a 影响的比例关系, 当 $a \in [0, 0.5)$, 表示虚拟机上的业务量降低, 相比本轮的带宽使用量, 前 k 次迭代拷贝期间网络带宽使用量要大, 则应该重点考虑历史带宽数据; 当 $a = 0.5$, 表示当前业务运行相对平稳; 当 $a \in (0.5, 1]$, 表示当前业务增大, 为防止以后业务量继续增加和抢占带宽现象的发生, 应该重点考虑本轮带宽使用量作为参考.

系数 a 会影响预留带宽值, 也会影响优化算法的可行性以及迁移的性能. 可以根据当前物理带宽使用量和物理带宽总量的比值作为系数 a 的值, 这是比较恰当的. a 的值可以真实反映当前物理网络资源的使用情况, 可以根据被迁移虚拟机产生的脏页率及时有效地给出本轮迭代过程中保证业务正常运行所需要的更合理的物理带宽.

3.3 带宽自适应算法

在初始阶段, 计算系统总带宽量、现阶段网络使用量及空闲的网络带宽. 迭代过程中, 根据迭代拷贝次数, 确定分配带宽的原则. 为首轮拷贝时, 根据初始化发送带宽即可. 当处于最后一轮停机拷贝时, 虚拟机将用上全部空闲带宽来传输剩余页面, 保证以最短时间将内存页面全部拷贝到目的虚拟机. 当迭代处于中间状态时, 需要计算预留带宽和修正预留带宽. 首先根据初始化阶段获取的计算机的物理网络信息计算带宽调整系数 a , 并根据脏页率和式(3)计算业务预留带宽. 然后, 对预留带宽进行修正. 业务的不断变化造成脏页率不断变化, 可能会造成计算得出的业务预留带宽大于现阶段的空闲的物理带宽. 当计算得出的业务预留带宽值大于现阶段的空闲带宽时, 需要调整业务预留带宽值, 即通过现阶段已使用的带宽量和计算得出的业务预留带宽值来对其进行调整. 最后, 计算出迭代拷贝需要的发送带宽, 并进入页面迭代拷贝发送阶段. 算法伪代码如下:

```

输入: 迭代轮次  $n$   输出: 本轮次分配带宽  $E_{bw}$ 
1 If( $n == 1$ ) //首轮迭代
2   Then return  $E_{bw}$  //初始化迭代发送带宽  $E_{bw}$ 
3 Else //物理网络总带宽  $T_{bw}$ , 使用带宽  $U_{bw}$ 
4   //计算物理空闲带宽  $F_{bw} = T_{bw} - U_{bw}$ 
5   if( $n == \text{last\_iter}$ ) //最后一轮迭代
6      $E_{bw} = F_{bw}$ 
7     Return  $F_{bw}$ 

```

```

8   Endif
9   计算带宽调整系数  $a = U_{bw} / T_{bw}$ 
10  获取本轮次脏页率  $U_{dirty}[n]$ 
11  // 计算前  $n-1$  轮脏页率的平均值
12   $D_{avg} = \sum_{i=1}^{n-1} U_{dirty_i} / (n-1)$ 
13  //计算业务当前预留带宽
14   $B_{bw} = (1-a) \times D_{avg} + a \times U_{dirty}[n]$ 
15  if( $B_{bw} > F_{bw}$ ) //修正业务预留带宽
16   $B_{bw} = F_{bw} \times (B_{bw} / (B_{bw} + U_{bw}))$ 
17  Endif
18  Return  $E_{bw} = F_{bw} - B_{bw}$ 
19  Endif

```

4 实验与结果分析

为了验证基于马尔科夫概率预测的基础上引入的带宽自适应算法(简记为 PAA)的性能, 将其与传统的 Pre-Copy(简记为 PCA)和基于概率预测的 Pre-Copy 迁移算法(简记为 PPC)进行对比分析.

4.1 实验环境

实验室选用三台配置为 Inter(R) Core(TM) i5-3210M 2.5 GHz、8 GB 内存、500 GB 硬盘的机器作为硬件平台, 同时采用 Ubuntu11.0 作为操作系统, Xen-4.1.0 作为虚拟化平台. 迁移不使用专用的网络, 与物理主机共享同一网络, 物理主机通过百兆交换机连接在同一局域网内.

4.2 压力测试

首先对改进的 PAA 机制进行压力测试, 测试的虚拟机内存大小为 128~2 048 MB, 主要测试了停机时间和总迁移时间.

(1) 停机时间

如图 2 所示, 在内存为 128 MB 和 256 MB 时, 三种策略机制的停机时间相差不大, 从 518 MB 开始, 本文的 PAA 优化机制可以根据业务量更好的调整迭代带宽, 使迭代效率更高, 在停机拷贝带宽为最大空闲带宽, 有效地减少了停机时间. 不同内存下的停机时间对比如图 2 所示.

(2) 总迁移时间

如图 3 所示, 在 128~256 MB 时, 三种算法总迁移时间相差不大. 从 512 MB 开始, PAA 就体现出了带宽自适应模块在迁移中的优势, 与 PPC 和 PCA 相比, 明显缩短了迁移总时间.

4.3 应用程序测试

实验采用三种工作负载进行测试, 第一种为空闲系统(idle, ID), 除了必须的系统程序外, 无额外的应用程序, 用来模拟低负载情况的迭代迁移. 第二种为较高负载系统(kernel-compile, KC), 即在系统

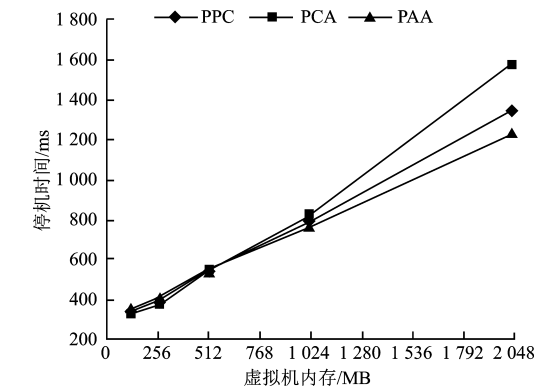


图 2 不同内存下停机时间对比图

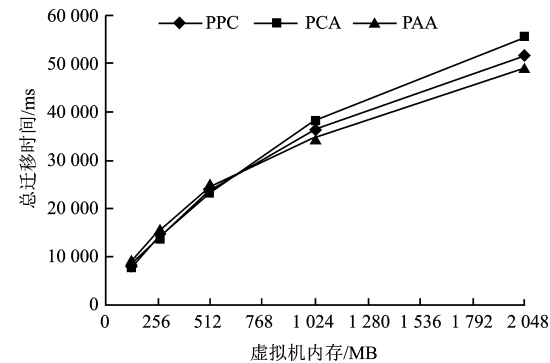


图 3 不同内存下的总迁移时间对比图

中编译 Linux 内核,来模拟较高脏页率的应用环境.第三种为间断性的内核编译(intermittent kernel-compilation,IKC),通过间断性的编译内核来模拟脏页率的突然急剧增加或者减少.实验所用虚拟机内存大小为 1 024 MB.

(1)停机时间

本文提出的 PAA 策略与 PCA 和 PPC 相比,停机时间有所减少,在 idle 应用环境下,PAA 并没有明显地缩短停机时间,该条件下,虚拟机中只运行了系统应用而无其他应用程序,对停机时间影响不大.而在内存较密集情况下,由于传送带宽会根据业务情况自动调整传输带宽,并在停机拷贝阶段,使用全部空闲带宽进行拷贝,缩短了停机时间.如图 4 所示,在 KC 负载下,PAA 相比 PPC 缩短了 9.16%;在 IKC 负载下,PAA 相比 PPC 缩短了 11.57%.在内存较密集情况下,能够较好的缩短停机时间.

(2)迁移总时间

如图 5 所示,PAA 相比 PCA 和 PPC 在迁移总时间上有明显地缩短,在 idle 情况下,PAA 并没有明显地缩短迁移时间,反而有少量的时间增长.但在较高脏页率和脏页率变化较大情况下,优化后的算法明显起到了缩短迁移总时间的作用.在 KC 负载下,PAA 相比 PPC 缩短了 9.45%;在 IKC 下,PAA

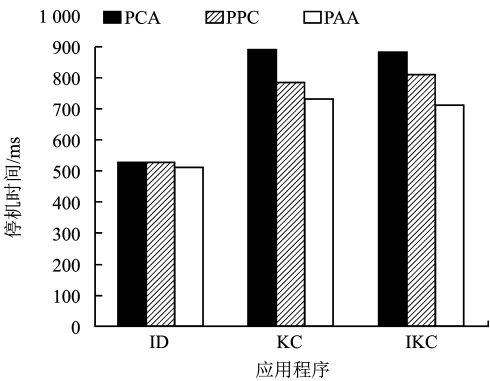


图 4 不同应用程序的停机时间对比图
相比 PPC 缩短了 11.46%.

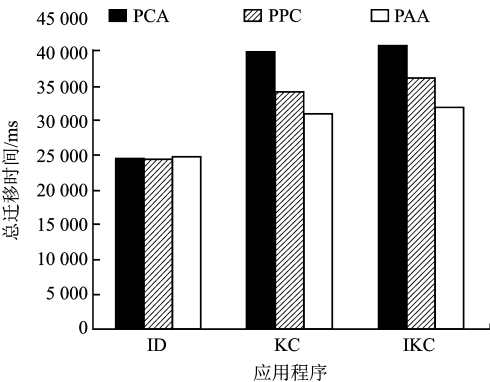


图 5 不同应用下的迁移总时间对比图
(3)迭代轮数

概率预测模块可以减少每轮传输的页面数量,带宽自适应模块可以在通过提高传输带宽来缩短每轮迭代所耗费的时间,使迭代收敛性^[8-9]更好.如图 6 所示,在脏页率低的情况下,由于 PAA 的概率预测和带宽自适应算法对资源的消耗,导致迭代轮数稍大于 PCA 和 PPA,当处于较高负载或者内存较密集时,由于带宽自适应模块,当脏页率较高的时自动提高带宽,缩短迁移时间,有利于降低脏页率,提高收敛速度,更快达到停机拷贝条件.

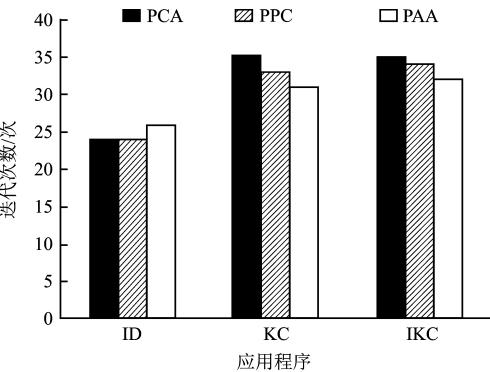


图 6 不同应用下的迭代轮数对比图

5 结束语

本文提出了一种基于概率预测的带宽自适应的预拷贝方法,并在 Xen 平台上进行实验,结果表明总迁移时间和停机时间及迭代轮数等都有所降低. 未来工作将针对复杂网络环境的迁移带宽分配算法进行优化以及对如何提高概率预测的可靠性进行研究.

参考文献:

[1] Nathuji R, Schwan K. Virtual power: Coordinated power management in virtualized enterprise system [C]//Proceedings of ACM Symposium on Operating Systems Principles (SOSP' 07). USA: IEEE, 2007: 265-278.

[2] Clark C, Fraser K, Hand S, et al. Live migration of virtual machines[C]//Proceedings of the 2nd Symposium on Networked Systems Design and Implementation NSDI' 05. Boston, USA: IEEE, 2005: 273-286.

[3] Wood T P, Shenoy P A, Venkataramani A, et al. Black-box and gray-box strategies for virtual machine migration[C]//Proceeding of the 4th USENIX Symposium on Networked Systems Design and Implementation (NSDI' 07). Cambridge, USA: IEEE, 2007: 229-242.

[4] 孙国飞,谷建华,胡金华,等. 基于预拷贝的虚拟机动态内存迁移机制改进[J]. 计算机工程, 2011, 32(10): 36-39.

[5] Sun Mingsong, Ren Wenwen. Improvement on dynamic migration technology of virtual machine based on xen[C]//Proceedings of the 8th International Forum on Strategic Technology. Washington DC USA: IEEE Press, 2013: 124-127.

[6] 阮敏. Xen 环境下实时迁移结构和算法研究[D]. 大连:大连海事大学, 2009.

[7] 刘诗海,孙宇清,刘古月. 面向业务特征的自适应虚拟机迁移带宽分配算法[J]. 计算机学报, 2013, 36(9): 1816-1825.

[8] Barhan P, Dragovic B, Fraser K, et al. Xen and the art of virtualization [C]//Proceedings of the 19 th ACM Symposium on Operating Systems Principles. New York, USA: ACM Press, 2003: 164-177.

[9] Sohan Rice R, Moore A, Hopper A W A. Predicting the performance of virtual machine migration[C]// Proceedings of the Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MAS-COTS). Florida, USA: IEEE, 2010: 37-46.

作者简介:

方义秋 女, (1963-), 副教授. 研究方向为云计算与软件工程.

陈玉洋(通讯作者) 男, (1989-), 硕士研究生. 研究方向为云计算与软件工程. E-mail: 1021535756@qq.com.

葛君伟 男, (1961-), 博士, 教授. 研究方向为云计算与软件工程.

(上接第 175 页)

[3] 董文瀚,马振强,解武杰,等. 非匹配不确定离散系统的无抖振积分滑模控制[J]. 控制与决策, 2015, 30(12): 2181-2186.

[4] Jing X J, Cheng L. An optimal PID control algorithm for training feedforward neural networks [J]. IEEE Transactions on Industrial Electronics, 2013, 60(6): 2273-2283.

[5] 章俊,张冬梅,王辛刚. 一类参数不确定连续正系统受限控制[J]. 信息与控制, 2015, 44(5): 592-597.

[6] 陆兴华,吴恩燊. 基于安卓客户端的智能家居电力控制优化设计[J]. 电力与能源, 2015, 35(5): 692-695.

[7] Feyzmahdavian H R, Charalambous T, Johansson M. Exponential stability of homogeneous positive systems of degree one with time-varying delays [J]. IEEE Transactions on Automatic Control, 2014, 59(6): 1594-1599

[8] Zhang J F, Han Z Z, Wu H. Robust finite-time stabil-

ity and stabilization of switched positive systems[J]. IET Control Theory and Applications, 2014, 8(1): 67-75.

[9] Ngoc P H A. Stability of positive differential systems with delay[J]. IEEE Transactions on Automatic Control, 2013, 58(1): 203-209.

作者简介:

张琰 女, (1981-), 硕士研究生, 讲师. 研究方向为物联网工程、多媒体技术. E-mail: 1286196111@qq.com

晏永红 女, (1969-), 硕士研究生, 副教授. 研究方向为电气自动化.

刘静 女, (1983-), 硕士研究生, 副教授. 研究方向为控制理论与控制工程.

周娅琴 女, (1980-), 硕士研究生, 工程师. 研究方向为计算机应用.