

私有云下基于相似度聚类的重删算法研究

马柄腾, 刘 丹

(电子科技大学 电子科学技术研究院, 四川 成都 610054)

摘 要: 随着虚拟化技术的日趋成熟与发展,越来越多的企业采用私有云平台来替代传统 PC 办公. 然而,虚拟磁盘数据的高度重复与超大容量导致了存储空间与系统能耗的巨大浪费. 为解决这一问题,首先提出了基于特征标签粗归类算法,将重删操作分散到云平台计算结点,有效地避免了传统重删算法的性能瓶颈;然后,提出了基于指纹 ID 相似度聚类算法,提高了各计算结点上的重删率;最后,通过实验对两种子算法进行分析,并验证了其有效性.

关键词: 私有云;主存储;oVirt;重复数据删除;相似度;聚类分析

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2017)09-0067-04

A Deduplication Algorithm Based on Similarity Clustering in Private Cloud

MA Bing-teng, LIU Dan

(Research Institute of Electronic Science and Technology, University of Electronic
Science and Technology of China, Chengdu 610054, China)

Abstract: With the maturity and development of virtualization technology, more and more enterprises adopt private cloud platform to replace the traditional PC of office. The highly duplicated and overlarge virtual disk data led to a huge waste of storage space and system power consumption. To solve this problem, first, we propose a classification algorithm based on feature tag. It re-distributes to deduplication the computing nodes of cloud platform, and effectively avoid the performance bottleneck of traditional deduplication algorithms. Then, it is proposed a similarity clustering algorithm based on fingerprint ID to improve the deduplication rate. Finally, through experiments, the two sub-algorithms is analyzed, and its effectiveness is verified.

Key words: private cloud; domain storage; oVirt; deduplication; similarity; cluster analysis

1 引言

云计算是当前信息技术领域的热门技术之一,作为云计算的一种重要形式,基础设施即服务 (Infrastructure as a Service, IaaS) 通过虚拟机向用户提供个性化和高质量的服务^[1]. 云平台提供服务的基础形式是虚拟机实例,其物理表现形式包括存储资源和计算资源 (CPU、内存与网络等). 为最大程度地利用存储资源,提高已有资源的利用率,将重复数据删除技术引入到存储系统,可以大规模地降低存储系统的存储空间开销^[2]. 云平台存储池优化主

要包括两个方面,分别是性能优化和空间优化. 国内外已经有不少学者对云平台存储池的性能优化问题进行了研究,但是对于云平台存储池空间优化的研究相对不足,而且很少考虑云平台性能问题或只面向备份容灾领域.

Ng 等人将 LiveDFS (基于重删技术的文件系统) 应用到开源云平台的虚拟机镜像存储系统中,减少了至少 40% 的存储空间^[3]. Zhang 等人提出了用户感知到重复数据删除算法,减少了 200% 以上的读磁盘操作,响应速度提升了 3 倍以上^[4].

当前的重删算法主要依据数据的时间局部性和

空间局部性来设计. 重删系统 I/O 处理流程分为 Chunking、ChunkHashing、FPindexing 和 DataStoring 等四个阶段. 在云平台主存储场景中 I/O 模型主要为随机的小颗粒 I/O, 这不仅会导致当前重删算法的性能下降, 同时对云平台的性能也会造成巨大的影响. 在重删过程中, 对性能影响最大的是指纹查询过程(FPindexing). 随着数据量的极速增加, 指纹的数量将远远超出内存阈值. 如果存储系统因内存资源不足而被迫采用磁盘索引, 则指纹查询会受到磁盘访问瓶颈的严重制约.

针对以上问题, 根据私有云架构以及虚拟磁盘相似性的特点, 采用分而治之的方法, 提出了一种基于相似度聚类重删算法.

2 基于相似度聚类重删算法

2.1 oVirt 云平台中虚拟磁盘特点

oVirt 是一个基于 x86 架构上的 KVM 虚拟化技术的开源 IaaS 云服务解决方案, 使用了 Node/Engine 分离结构. Engine 在整个系统中充当管理者的角色并对外提供管理服务, Node 上运行 VDSM, 实现网络、存储器、虚拟机的创建与修改的功能. 为了简化管理, oVirt 中抽象出了数据中心(Data Center)概念, 在数据中心中, 使用一个存储池(Storage Pool)来保存虚拟磁盘数据(Virtual Data).

oVirt 云平台的设备模拟主要通过 KVM 内核模块和 QEMU 实现. KVM 是一种基于硬件虚拟化技术的虚拟机监视器(Virtual Machine Monitor, 又称为 Hypervisor), 负责为每一台虚拟机实例分配适量的内存、CPU、网络和硬盘, 并加载所有虚拟机的客户操作系统. 轻量型的 QEMU 进程负责处理客户操作系统的 I/O, 并为每个客户机操作系统提供若干模拟硬件设备(如磁盘、网卡、BIOS 等). 客户虚拟机的磁盘是通过 QEMU 在宿主机上模拟的虚拟磁盘, 在存储池中只是一个物理磁盘系统上的一个文件. 因此, 虚拟磁盘的读写操作会映射到虚拟磁盘文件在物理文件系统上的读写操作.

图 1 描述了客户虚拟机的磁盘操作控制流. 客户虚拟机的一个磁盘 I/O 请求引发一次陷入, 然后 KVM 通知 QEMU 用户空间进程负责处理. 所有虚拟机的磁盘数据缓存首先会保存在宿主机的数据缓存内, 每一个虚拟机通过主机页面缓存获取数据并保留在虚拟机的数据缓存中. 根据虚拟机的磁盘操作控制流可知, 虚拟磁盘的 I/O 操作的处理不需要修改客户操作系统, 即对于客户操作系统而言虚拟

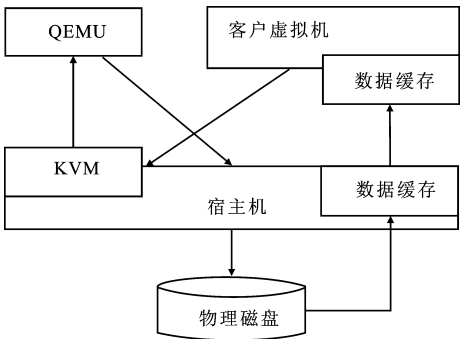


图 1 虚拟机的磁盘操作控制流

磁盘 I/O 流的处理是透明的.

2.2 重复数据删除技术基本原理

重复数据删除技术也称为智能压缩技术或单一实例存储技术. 其工作流程是自动搜索重复数据, 为相同数据保留唯一副本, 并使用指向单一副本的指针替换掉其他重复副本. 该技术可以消除冗余数据、降低存储容量的存储成本, 提高已有资源利用率. 重复数据删除技术的基本实现有三步: (1) 在存储数据时将数据分块; (2) 对每个划分好的数据块均通过 hash 算法得到指纹; (3) 比较已存储的数据中是否已经包含该数据, 未包含则存储该数据, 否则过滤掉这部分数据的备份, 并以指向唯一实例的指针取代. 如图 2 所示.

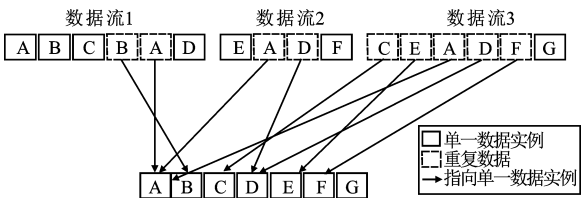


图 2 重复数据删除技术

2.3 虚拟磁盘的相似度分析

2.3.1 基于特征标签粗归类算法

虚拟机实例在逻辑上是一个完全独立的虚拟磁盘, 对应于物理磁盘上的一个单独的文件. 虚拟机实例具有一些固有的特征标签(又称为特征属性), 而拥有相同特征标签的虚拟磁盘具有一定程度的数据重复, 这也是云平台数据中心存在 80% 以上数据冗余的主要原因. 根据云平台的使用情况, 这些特征标签主要分为两类, 包括物理特征标签和关系特征标签.

虚拟磁盘的物理标签主要由数据内容决定, 主要包括操作系统、应用程序、数据文件等. Li 等人对虚拟磁盘镜像进行分析, 得到不同操作系统和不同文件格式的重复率, 分析结果如表 1 所示^[5].

表 1 不同操作系统的数据重复率 %

类别	Fedora 8	Fedora 9	Ubuntu 8.04	Open SUSE 11	Windows XP(FAT32)
Fedora 7	34	22	8	15	—
Fedora 8		31	10	16	—
Fedora 9			11	21	—
Ubuntu 8.04				8	—
Windows XP(NTFS)					27

利用云平台计算结点对聚类后虚拟磁盘进行重复数据删除,可以大规模缩减指纹数量,并达到内存级指针索引查询.同时,对云平台数据中心进行聚类分析的重删处理会加剧云平台的资源竞争.本文提出一种基于特征标签粗归类算法,将用户创建的虚拟磁盘采用 Affinity Propagation(AP)算法^[6]进行归类,从而达到降低计算维度,并且保证一定的重删率.

定义 1 设两个虚拟磁盘分别为 f 和 g , f 的标签属性为 $\{i_1, i_2, \dots, i_m\}$, g 的标签属性为 $\{j_1, j_2, \dots, j_n\}$,并根据经验构建一个标签属性的重复数据概率表,记录两个标签的重复数据概率,将所有虚拟磁盘都作为潜在的归类中心,记为 Exp ,用 $s(i, k)$ 来表示 f 与 Exp 之间的特征标签相似度.这些相似度组成 $N \times N$ 的相似度矩阵 S .

以 S 矩阵的对角线上的数值 $s(k, k)$ 作为 k 点能否成为聚类中心的评判标准,将该评判标准称为参考度 p (Preference). AP 算法中两种消息为 responsibility 和 availability. $r(i, k)$ 表示从点 i 发送到候选聚类中心 k 的数值信息,该数值信息反映 k 点是否适合作为 i 点的聚类中心. $a(i, k)$ 则从候选聚类中心 k 发送到 i 的数值消息,反映点 i 是否选择 k 作为其聚类中心. $r(i, k)$ 和 $a(i, k)$ 越强,则 k 点作为聚类中心的可能性就越大,并且 i 隶属于 k 点为聚类中心的聚类的可能性也越大^[6]. AP 算法通过迭代过程不断更新每一个点的吸引度和归属感,直到产生 m 个高质量的 Exp ,同时将其余的数据点分配到相应的聚类中.最后再将不同归类转移到不同的计算结点,并完成重复数据删除过程.

2.3.2 基于指纹 ID 相似度聚类算法

采取 4k 大小的固定长度分块算法对云平台下的虚拟磁盘进行分块,并构建指纹数据库.虚拟磁盘的指纹数据库包含了每一个虚拟机实例的所有虚拟磁盘指纹信息,一个磁盘的指纹信息是该虚拟磁盘数据的唯一准值标识.

定义 2 设两个虚拟磁盘分别为 f 和 g , 磁盘指纹 ID 对应的集合分别为 $S_f = \{block.id \mid block \in f\}$ 和 $S_g = \{block.id \mid block \in g\}$,两个虚拟磁盘相同的数据块指纹数为 $FPMin(f, g) = |S_f \cap S_g|$,数据块指纹总数目为

$$FPMax(f, g) = |S_f| + |S_g| - FPMin(f, g) \tag{1}$$

那么,磁盘 f 和 g 的相似度为

$$VDSim(f) = \frac{FPMin(f, g)}{FPMax(f, g)} \tag{2}$$

虚拟磁盘相似度的值域区间为 $[0, 1]$, 如果值超过了临界值,即可判定两个虚拟磁盘相似.

定义 3 设 n 个虚拟磁盘组成的集合为簇 $F = \{f_1, f_2, \dots, f_n\}$, 数据中心 $\Omega = F_1 \cup F_2 \dots \cup F_m$, 虚拟磁盘 f , 定义虚拟磁盘 f 与簇 F 的相似度为

$$VDCSim(F) = \frac{FPMin(f, F)}{FPMax(f, F)} \tag{3}$$

虚拟磁盘与簇之间相似度的值域区间为 $[0, 1]$, 其大小代表了该虚拟磁盘与簇的相似度,定义与虚拟磁盘 f 相似度最大的簇为虚拟磁盘聚类阈值为

$$MaxSim = \{F_i \mid \max_i VDCSim(F_i)\} \tag{4}$$

根据虚拟磁盘聚类阈值,可以找出与该虚拟磁盘相似度最高的簇,从而达到获得最佳的重删率.

定义 4 一个簇中的虚拟磁盘进行重复数据删除操作,并产生相应的指纹库,虚拟磁盘 f 的指纹集合 S_f 可以通过抽样的方式进行简化,故定义数据指纹的热度为同一块数据块的被引用次数.将数据块指纹的热度记录在指纹元数据中,抽样采取规则根据数据块指纹的热度选取指纹热度超过一定阈值的指纹 ID 进行统计.

经过基于特征标签粗归类算法的处理,云平台数据中心的重删操作已经从全局性转变为各个计算结点上的局部性操作,实现了内存级指纹搜索查询,从而使重删算法的性能得到优化.通过计算虚拟磁盘聚类阈值可以找到其最佳的聚类分组,从而获得更好的重删率.

基于指纹 ID 相似度聚类算法过程如下:

初始化 N 个簇,根据数据块指纹热度进行抽样,统计每个簇内的虚拟磁盘的指纹集合.

选择待聚类的虚拟磁盘,分别计算其与 N 个簇的相似度 $VDCSim(F)$.

比较得到 $VDCSim(F)$, 获的最值,即虚拟磁盘聚类阈值 $MaxSim$.

根据虚拟磁盘聚类阈值,在云平台空闲时将待聚类的虚拟磁盘转移到最佳的计算结点。

重复第二步,直到处理完所有虚拟磁盘。

由于云平台虚拟磁盘是动态增加,并不断变化的,所以需要结合云平台性能均衡技术,不定时的进行基于指纹 ID 相似度聚类的重删操作,从而使得云平台时刻保持最佳的系统性能与重删率。

3 实验分析

本文实验环境是 oVirt 云平台,环境配置在 5 台服务器,其中包括 1 台 Engine 结点、3 台 Node 结点和 1 台备份服务器。每台物理服务器内存 64 GB, CPU24 核,硬盘 1TB。所有设备通过千兆网络设备进行连接。

3.1 基于特征标签粗归类算法

云平台数据中心属于主存储领域,其对于系统响应以及系统性能要求极高。对于已有未具备重删功能的数据中心或备份虚拟磁盘库,以及用户逐渐添加的新虚拟磁盘,实时性的进行基于相似度聚类的重删操作会严重影响用户体验。因此,本文采取了一种基于虚拟机磁盘特征标签粗归类算法,从而满足云平台的实时性需求以及一定重删度的去冗余操作。不同于以往的虚拟机仅根据性能均衡来生成新的虚拟磁盘,粗归类性能分析成为了需要我们关注的一个重要元素。

实验针对未具备重删功能的备份虚拟磁盘库进行恢复云平台的数据中心,备份的虚拟磁盘数量为 30 个,比较采取粗归类与裸恢复的时间,实验结果如图 3 所示。

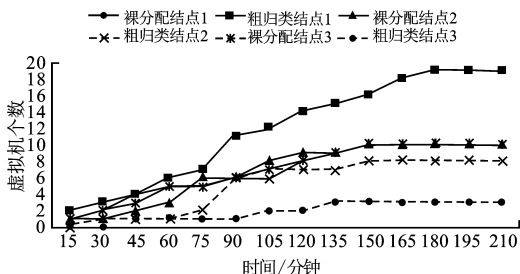


图3 裸恢复与粗归类恢复时间比较

上述实验分析可知,裸恢复的虚拟磁盘分布较为均匀,且最终各结点的虚拟磁盘个数基本相同,而粗恢复由于基于特征标签粗归类算法,其分布情况则明显分布不均。由于归类算法需要一定的时间,所以粗恢复完成的时间也较长。

3.2 基于指纹 ID 相似度聚类算法

在进行粗恢复后的各计算结点上的重删率由于采

用特征标签作为相似度分析的标准,因此其重删率较低。而基于指纹 ID 相似度聚类的重删算法采用指纹 ID 的集合作为相似度分析的标准,可以唯一标识一个虚拟磁盘,从而在各计算结点都达到最佳的重删率。

在上述备份虚拟磁盘恢复实验的基础上,我们进行一轮基于指纹 ID 相似度聚类算法的重删操作,实验结果如图 4 所示。结点一与结点二的重删率大幅度提升,这也证明了基于指纹 ID 相似度聚类算法可以获得比基于特征标签粗归类算法更佳的重删效果,从而弥补了粗粒度归类的缺点。对于结点三,两种情况下重删率相同,可以位于结点三上的基于特征标签粗归类在特定情况下可以达到基于指纹 ID 相似度聚类的效果,从侧面也证明了基于特征标签粗归类算法的有效性。

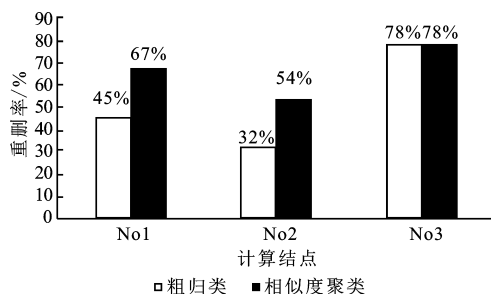


图4 粗归类与相似度聚类的重删率比较

4 结束语

在企业级私有云的快速发展下,云平台主存储下基于重删算法的空间优化能够获得更大的经济效益。然而,常规的重删算法往往不能满足云平台的数据中心的大规模与高效实时性。为了能够在私有云云平台主存储系统中应用重删算法,我们结合 oVirt 私有云的特点,利用各计算结点的性能,并将机器学习中的相似度聚类分析引入聚类算法中,从而将重删系统由全局性转向局部性,并且可以获得最佳重删率。本文提出的基于虚拟磁盘特征标签粗归类算法可以对虚拟磁盘进行粗粒度的归类,并通过虚拟机指纹 ID 集合决定的相似度聚类算法,进一步提高了云平台中各个计算结点的重删率。

参考文献:

- [1] WANG Jiajun, LV Zhihui, WU Jie, et al. Cloud computing technology development analysis and applications discussion[J]. Computer Engineering and Design, 2010, 31(20): 4404-4408.

(下转第 76 页)

了这两种方法的特点,然后在传统结构上设计了电流仅为 264 nA 的低功耗带隙,电源电压为 1.8 V 输出电压为 1.196 4 V,温度系数仅为 7 ppm/°C,电源抑制比 1K 时为-78 dB,通过一些方法和措施将工艺对电路的影响降到最低,从而提高了电路的精度.设计过程中仿真了器件所有 corner,温度从-40~125°C 和电源电压±10%的情况、电源电压为 1.8 V 的时候最恶劣,corner 总电流为 343 nA,版图面积为 105 $\mu\text{m}\times 110 \mu\text{m}$.

本文很好地实现了低功耗高精度带隙在同等面积上功耗最低,在同等功耗下面积最小,温度系数较小,PSRR 较高,误差较小,电压较稳定的特点,实现了在精度、功耗面积之间的折中,很好地满足了设计要求和工业的需要.

参考文献:

- [1] Ken Ueno, Tetsuya Hirose, Tetsuya Asai, et al. A 300nW, 15 ppm/°C, 20 ppm/V CMOS voltage reference Circuit consisting of subthreshold MOSFETs[J]. IEEE Journal of Solid-State Circuits, 2009,44:2047-2054.
- [2] Yuji Osaki, Tetsuya Hirose, Nobutaka Kuroki, et al. 1.2-V supply, 100-nW, 1.09-V bandgap and 0.7-V supply, 52.5-nW, 0.55-V subbandgap reference circuits for nanowatt CMOS LSIs[J]. IEEE Journal of Solid-State Circuits, 2013,48:1530-1538.
- [3] Cheng M H, Wu Z W. Low-power low-voltage reference using peaking current mirror circuit Electron[J]. Lett., 2005, 41(10):572-573.
- [4] Banba H, Shiga H, Umezawa A, et al. A CMOS

bandgap reference circuit with sub-1-V operation[J]. IEEE Journal of Solid-State Circuits, 1999,34: 670-674.

- [5] 何乐年,王忆. 模拟集成电路设计与仿真 [M]. 北京: 科学出版社,2008.
- [6] RAZAVI B. 模拟 CMOS 集成电路设计 [M]. 陈贵灿,程军,译. 西安:西安交通大学出版社,2002.
- [7] Wang A, Clhoun B H, Chandracasan A P. Sub-threshold design for ultra low-Power systems [M]. New York: Springer, 2006.
- [8] Taur Y, Ning T H. Fundamentals of modern VLSI devices [M]. Cambridge, U. K.: Cambridge Univ. Press, 2002.
- [9] Zhang Hualei, Xiao Zhibin, Xi Tan, et al. Low-power sub-1-V compact bandgap reference for passive RFID tags[J]. Electron. Lett., 2015,51(11):816.
- [10] Filippo Neri, Thomas Brauner, Eric De Mey, et al. Low-power, wide supply voltage bandgap reference circuit in 28nm CMOS[C]// IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). IEEE, 2015:1-6.

作者简介:

纪书江 男,(1989-),硕士研究生.研究方向为模拟集成电路设计. E-mail:jishujiang@ime.ac.cn.

霍长兴 男(满族),(1984-),研究实习员.研究方向为模拟集成电路设计.

孙海涛 男,(1987-),研究实习员.研究方向为存储器定制电路设计.

刘璟 男,(1985-),助理研究员.研究方向为 Flash 存储器电路设计及可靠性.

(上接第 70 页)

- [2] Clements A, Ahmad I, Vilayannur M. et al. Decentralized deduplication in SAN cluster file systems[C] // Proc of the USENIX ATC'09. Berkeley,CA: USENIX, 2009:101-114.
- [3] Ng C H, Patrick P. RevDedup: A reverse deduplication storage system optimized for reads to latest backups. //Proc. Of the 4th ACM SIGOPS Asia-Pacific Workshop on Systems. Singapore, 2013. [doi: 10.1145/2500727.2500731].
- [4] Zhang H Y, Zhou J C, Zhou Yibo, et al. User-Aware De-Duplication Algorithm[J]. Journal of Software, 2015,26(10):2581-2595.

- [5] Li Gouri A, Hensbergen E. Experiences with content addressable storage and virtual disks // Proc of USENIX WIOV'08. Berkeley: USENIX, 2008: 5-10.

- [6] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(16): 972-976.

作者简介:

马柄腾 男,(1990-),硕士.研究方向为分布式存储、云计算、网络安全. E-mail:mabingteng@sina.com.

刘丹 男,(1969-),博士,副教授.研究方向为网络安全、分布式并行计算、云计算与大数据处理.