



(21) 申请号 201810620396.8

(22) 申请日 2018.06.15

(65) 同一申请的已公布的文献号
申请公布号 CN 109284068 A

(43) 申请公布日 2019.01.29

(30) 优先权数据
62/534,647 2017.07.19 US
15/887,980 2018.02.02 US

(73) 专利权人 三星电子株式会社
地址 韩国京畿道水原市灵通区三星路129号

(72) 发明人 杨征宇 汤玛斯大卫·艾文斯
艾伦·安德鲁斯 克雷·迈尔斯
汤玛斯罗里·伯尔特

(74) 专利代理机构 华进联合专利商标代理有限公司 44224

专利代理师 刘培培 黄隶凡

(51) Int.Cl.

G06F 3/06 (2006.01)

(56) 对比文件

CN 103959254 A, 2014.07.30
US 9342253 B1, 2016.05.17
US 2015295791 A1, 2015.10.15
WO 2005017737 A2, 2005.02.24
CN 104516768 A, 2015.04.15
US 9116914 B1, 2015.08.25
US 2011197044 A1, 2011.08.11
US 2010115049 A1, 2010.05.06
CN 104657215 A, 2015.05.27
CN 104298339 A, 2015.01.21
US 2016085481 A1, 2016.03.24

审查员 韩洁

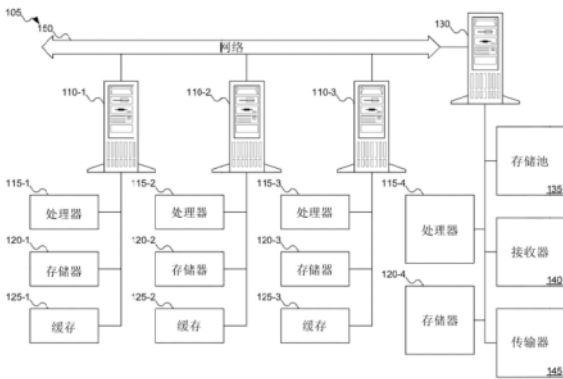
权利要求书4页 说明书29页 附图13页

(54) 发明名称

数据存储管理系统、方法及装置

(57) 摘要

本发明提供一种数据存储管理系统、数据存储管理方法及数据存储管理装置。本发明公开一种用于执行自动分层的系统。系统可以包含提供多个资源并组织成存储层的多个存储装置。存储装置能够存储虚拟机的数据。接收器能够接收虚拟机的输入/输出命令和性能数据。传输器能够传输对于输入/输出命令的响应。自动分层控制器能够选择存储层以存储虚拟机的数据并且能够响应于性能数据而在存储层之间迁移数据。将虚拟机的性能改变和在存储层之间迁移数据的迁移成本作为考虑因素,存储层的选择能够优化所有虚拟机在所有存储层中的性能。



1. 一种数据存储管理系统,包括:

提供多个资源的多个存储装置,所述多个存储装置被组织成多个存储层并且存储第一虚拟机的第一数据和第二虚拟机的第二数据;

接收器,接收来自所述第一虚拟机的第一输入/输出命令、来自所述第二虚拟机的第二输入/输出命令、模拟所述第一虚拟机在所述多个存储层中的性能的第一性能数据,以及模拟所述第二虚拟机在所述多个存储层中的性能的第二性能数据;

传输器,将对于所述第一输入/输出命令的第一响应传输到所述第一虚拟机,并且将对于所述第二输入/输出命令的第二响应传输到所述第二虚拟机;以及

自动分层控制器,选择第一存储层以存储所述第一虚拟机的所述第一数据,选择第二存储层以存储所述第二虚拟机的所述第二数据,并且响应于所述第一性能数据和所述第二性能数据来迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层,

其中所述自动分层控制器更用以选择所述第一存储层以存储所述第一虚拟机的所述第一数据并且选择所述第二存储层以存储所述第二虚拟机的所述第二数据,从而优化所有虚拟机在所述多个存储层中的性能,

其中所述自动分层控制器更用以将由迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层所引起的性能改变和迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层的迁移成本作为考虑因素,

其中所述自动分层控制器包含使用所述第一性能数据和所述第二性能数据生成资源矩阵的资源矩阵生成器,所述资源矩阵使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联。

2. 根据权利要求1所述的数据存储管理系统,其中所述自动分层控制器包含:

层选择模块,响应于所述第一性能数据和所述第二性能数据选择所述第一虚拟机的所述第一存储层和所述第二虚拟机的所述第二存储层;以及

迁移模块,迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层。

3. 根据权利要求1所述的数据存储管理系统,其中所述资源矩阵生成器包含度量转换器,其根据所述多个资源的第二资源将所述第一虚拟机在所述多个存储层中的每一个中的第一性能度量转换成所述第一虚拟机在所述多个存储层中的每一个中的第二性能度量。

4. 根据权利要求3所述的数据存储管理系统,其中所述第一性能度量包含所述第一虚拟机的第一平均输入/输出延迟数据。

5. 根据权利要求1所述的数据存储管理系统,其中:

所述自动分层控制器包含线性回归模块,其对所述第一性能数据执行第一线性回归并且对所述第二性能数据执行第二线性回归;以及

所述自动分层控制器能够操作以响应于所述第一线性回归和所述第二线性回归而选择所述第一存储层以存储所述第一虚拟机的所述第一数据且选择所述第二存储层以存储所述第二虚拟机的所述第二数据。

6. 根据权利要求5所述的数据存储管理系统,其中所述自动分层控制器能够操作以响

应于所述第一线性回归和所述第二线性回归,使用所述第一虚拟机和所述第二虚拟机的估计平均输入/输出延迟来选择所述第一存储层以存储所述第一虚拟机的所述第一数据并且选择所述第二存储层以存储所述第二虚拟机的所述第二数据。

7. 根据权利要求1所述的数据存储管理系统,其中所述自动分层控制器更用以选择所述第一存储层来存储所述第一虚拟机的所述第一数据,选择所述第二存储层来存储所述第二虚拟机的所述第二数据,并且每个迁移期一次响应于所述第一性能数据和所述第二性能数据迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层。

8. 根据权利要求7所述的数据存储管理系统,其中:

所述第一性能数据包含来自所述迁移期内的多个监测期的第一性能数据;以及

所述第二性能数据包含来自所述迁移期内的所述多个监测期的第二性能数据。

9. 根据权利要求1所述的数据存储管理系统,进一步包括:

第一虚拟机服务器,存储所述第一虚拟机;以及

第二虚拟机服务器,存储所述第二虚拟机。

10. 根据权利要求9所述的数据存储管理系统,其中所述第一虚拟机服务器包含生成所述第一性能数据的输入/输出滤波器,所述第一性能数据模拟所述第一虚拟机在所述多个存储层中的性能。

11. 根据权利要求10所述的数据存储管理系统,其中所述输入/输出滤波器能够操作以将延迟增加到来自所述第一虚拟机的多个输入/输出命令中,所述多个输入/输出命令包含所述第一输入/输出命令,并且响应于所增加的延迟确定所述第一虚拟机的平均输入/输出延迟。

12. 一种数据存储管理方法,包括:

接收模拟第一虚拟机在多个存储层中的性能的第一性能数据;

接收模拟第二虚拟机在所述多个存储层中的性能的第二性能数据;

使用所述第一性能数据和所述第二性能数据生成资源矩阵,所述资源矩阵使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联;

使用所述资源矩阵选择所述多个存储层中的第一存储层以存储所述第一虚拟机的第一数据;

使用所述资源矩阵选择所述多个存储层中的第二存储层以存储所述第二虚拟机的第二数据;以及

迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层,

其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据使所有虚拟机在所述多个存储层中的性能优化,

其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将由迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层所引起的性

能改变作为考虑因素,以及

其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层的迁移成本作为考虑因素。

13. 根据权利要求12所述的数据存储管理方法,其中:

使用所述资源矩阵选择所述多个存储层中的所述第一存储层以存储所述第一虚拟机的所述第一数据的步骤包含使用所述资源矩阵和所述多个存储层中的每一个的所述专业化信息选择所述多个存储层中的所述第一存储层以存储所述第一虚拟机的所述第一数据;以及

使用所述资源矩阵选择所述多个存储层中的所述第二存储层以存储所述第二虚拟机的所述第二数据的步骤包含使用所述资源矩阵和所述多个存储层中的每一个的所述专业化信息选择所述多个存储层中的所述第二存储层以存储所述第二虚拟机的所述第二数据。

14. 根据权利要求13所述的数据存储管理方法,其中所述多个存储层中的每一个的所述专业化信息确认所述多个存储层中的每一个是否支持所述多个资源中的每一个。

15. 根据权利要求14所述的数据存储管理方法,其中模拟所述第一虚拟机在所述多个存储层中的性能的所述第一性能数据包含使所述第一虚拟机的平均输入/输出延迟与所增加的延迟相关联的多个数据点。

16. 根据权利要求12所述的数据存储管理方法,其中接收模拟所述第一虚拟机在所述多个存储层中的性能的所述第一性能数据包含:

在输入/输出滤波器处接收来自所述虚拟机的多个第一输入/输出命令;

将第一延迟增加到所述多个第一输入/输出命令中的每一个中;以及

针对所述多个第一输入/输出命令计算所述虚拟机的第一平均输入/输出延迟。

17. 根据权利要求12所述的数据存储管理方法,其中迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层的迁移成本包含迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层所需要的迁移时间。

18. 一种数据存储管理装置,包括非暂时性存储介质,所述非暂时性存储介质上存储有指令,所述指令由机器执行时引起以下操作:

接收模拟第一虚拟机在多个存储层中的性能的第一性能数据;

接收模拟第二虚拟机在所述多个存储层中的性能的第二性能数据;

使用所述第一性能数据和所述第二性能数据生成资源矩阵,所述资源矩阵使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联;

使用所述资源矩阵选择所述多个存储层中的第一存储层以存储所述第一虚拟机的第一数据;

使用所述资源矩阵选择所述多个存储层中的第二存储层以存储所述第二虚拟机的第二数据;以及

迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层,

其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据使所有虚拟机在所述多个存储层中的性能优化,

其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将由迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层所引起的性能改变作为考虑因素,以及

其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将迁移所述第一虚拟机的所述第一数据到所述第一存储层及/或迁移所述第二虚拟机的所述第二数据到所述第二存储层的迁移成本作为考虑因素。

19. 根据权利要求18所述的数据存储管理装置,其中接收模拟第一虚拟机在多个存储层中的性能的第一性能数据包含:

在输入/输出滤波器处接收来自所述虚拟机的第一输入/输出命令;

将第一延迟增加到所述第一输入/输出命令中;以及

针对所述第一输入/输出命令计算所述虚拟机的第一平均输入/输出延迟。

数据存储管理系统、方法及装置

技术领域

[0001] 本发明概念大体上涉及数据存储,且更具体地说,涉及将虚拟机数据存储在哪个存储层中的自动管理。

背景技术

[0002] 常规的数据存储中心依赖于硬盘驱动器和其它早期形式的数据存储。由于硬盘驱动器和其它早期形式的数据存储具有相对较高的延迟,所以虚拟机缓存频繁存取本地存储装置中的数据。通常,缓存数据的相关益处为数据存取时间增大十倍或大于十倍。

[0003] 但是随着快闪存储变得更常用且更负担得起,这种在数据存储中心与缓存之间的存取时间上的差异减小。这种缓存的益处减少意味着数据频繁地专门存储在数据中心中且根本不缓存。当结合将数据的冗余复本(缓存和数据中心两者中的)存储在快闪存储装置中是昂贵的事实时,以最高效的方式将数据存储在与数据存储中心中变得更重要。

[0004] 另外,不同于常规的存储装置,例如硬盘驱动器,快闪存储可以具有不同的专业化。由于常规的存储装置并不考虑装置专业化,所以使用常规技术进行数据存储的任何层选择是次优的。

[0005] 仍需要一种使用快闪存储更优化地将数据存储在与数据存储中心中的方式,所述快闪存储同时提供令人满意的存取时间且利用任何由不同存储层提供的专业化。

发明内容

[0006] 本发明概念的实施例包含一种系统,其包括:提供多个资源的多个存储装置,所述多个存储装置组织成多个存储层并且存储第一虚拟机的第一数据和第二虚拟机的第二数据;接收器,接收来自所述第一虚拟机的第一输入/输出(I/O)命令、来自所述第二虚拟机的第二I/O命令、模拟所述第一虚拟机在所述多个存储层中的性能的第一性能数据,以及模拟所述多个存储层中的所述第二虚拟机的性能的第二性能数据;传输器,将对于所述第一I/O命令的第一响应传输到所述第一虚拟机并且将对于所述第二I/O命令的第二响应传输到所述第二虚拟机;以及自动分层控制器,选择第一存储层以存储所述第一虚拟机的所述第一数据、选择第二存储层以存储所述第二虚拟机的所述第二数据,且响应于所述第一性能数据和所述第二性能数据迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个,其中所述自动分层控制器可操作以选择所述第一存储层以存储所述第一虚拟机的所述第一数据并且选择所述第二存储层以存储所述第二虚拟机的所述第二数据来优化所有虚拟机在所述多个存储层中的性能,其中所述自动分层控制器可操作以将由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所引起的性能改变和迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个的迁移成本作为考虑因素。

[0007] 本发明概念的实施例包含一种方法,其包括:接收模拟第一虚拟机在多个存储层

中的性能的第一性能数据、接收模拟第二虚拟机在所述多个存储层中的性能的第二性能数据、使用所述第一性能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵、使用所述资源矩阵选择所述多个存储层中的第一存储层以存储所述第一虚拟机的第一数据、使用所述资源矩阵选择所述多个存储层中的第二存储层以存储所述第二虚拟机的第二数据,以及迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个;其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据使所有虚拟机在所述多个存储层中的性能优化,其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所引起的性能改变作为考虑因素,以及其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个的迁移成本作为考虑因素。

附图说明

- [0008] 图1示出根据本发明概念的实施例的可以托管虚拟机并且使用经装备以执行自动分层的存储池的连网机器;
- [0009] 图2示出图1的机器的额外细节;
- [0010] 图3示出根据本发明概念的实施例的在自动分层的情况下与图1的存储池进行交互的图1的一个服务器的额外细节;
- [0011] 图4示出由不同存储装置提供的不同可用资源;
- [0012] 图5示出将从一个存储装置迁移到另一个存储装置的图3的虚拟机;
- [0013] 图6A到图6B示出在不同时间点处虚拟机到存储层的最优分配;
- [0014] 图7示出增加延迟并且确定图3的虚拟机的性能数据的图3的I/O滤波器;
- [0015] 图8示出图3的自动分层控制器的细节;
- [0016] 图9示出图7的性能数据转换成使虚拟机、存储层和所述存储层的资源相关联的矩阵;
- [0017] 图10示出图7的性能数据的线性回归;
- [0018] 图11示出关于图3的存储层的专业化信息;
- [0019] 图12示出待用于将图3的虚拟机迁移到图3的不同存储层中的得分矩阵的计算;
- [0020] 图13示出监测期与迁移期之间的关系;
- [0021] 图14示出根据本发明概念的实施例的用于图3的I/O滤波器将延迟增加到输入/输出(I/O)命令以产生图7的性能数据的实例过程的流程图;
- [0022] 图15A到图15B示出根据本发明概念的实施例的用于图3的自动分层控制器管理虚拟机在存储层中的迁移的实例过程的流程图;
- [0023] 图16示出根据本发明概念的实施例的用于图3的自动分层控制器生成图9的资源

矩阵的实例过程的流程图。

- [0024] 附图标号说明
- [0025] 105:系统;
- [0026] 110:虚拟机服务器;
- [0027] 110-1、110-2、110-3:虚拟机服务器;
- [0028] 115:处理器;
- [0029] 115-1、115-2、115-3、115-4:处理器;
- [0030] 120:存储器;
- [0031] 120-1、120-2、120-3、120-4:存储器;
- [0032] 125-1、125-2、125-3:缓存;
- [0033] 130:数据中心服务器;
- [0034] 135:存储池;
- [0035] 140:接收器;
- [0036] 145:传输器;
- [0037] 205:存储器控制器;
- [0038] 210:时钟;
- [0039] 215:网络连接器;
- [0040] 220:总线;
- [0041] 225:用户接口;
- [0042] 230:输入/输出引擎;
- [0043] 305-1、305-2、305-3:虚拟机;
- [0044] 310:应用程序;
- [0045] 315:操作系统;
- [0046] 320:管理程序;
- [0047] 325:I/O滤波器;
- [0048] 330:自动分层后台程序;
- [0049] 335:自动分层控制器;
- [0050] 340-1、340-2、340-3:存储层;
- [0051] 345:存储装置;
- [0052] 405:资源;
- [0053] 410:吞吐量;
- [0054] 415:带宽;
- [0055] 420:容量;
- [0056] 425:写入放大因子;
- [0057] 505:虚拟机文件;
- [0058] 510:虚拟机文件;
- [0059] 515:箭头;
- [0060] 705:I/O命令;
- [0061] 710:延迟;

- [0062] 715:响应;
- [0063] 720:性能数据;
- [0064] 805:线性回归模块;
- [0065] 810:资源矩阵生成器;
- [0066] 815:度量转换器;
- [0067] 820:专业化信息;
- [0068] 825:层选择模块;
- [0069] 830:迁移模块;
- [0070] 905:资源矩阵;
- [0071] 910:二维矩阵;
- [0072] 915:二维矩阵;
- [0073] 920:二维矩阵;
- [0074] 1005:线性回归;
- [0075] 1010:线性回归;
- [0076] 1205:得分矩阵;
- [0077] 1305:迁移期;
- [0078] 1310:监测期;
- [0079] 1405、1410、1415、1418、1420、1425、1430、1505、1510、1515、1520、1525、1530、1535、1605、1610、1615、1620:块;
- [0080] 1625:虚线;
- [0081] 3615:性能数据;
- [0082] T0、T1:时间。

具体实施方式

[0083] 现在将详细参考本发明概念的实施例,在附图中图示所述实施例的实例。在以下详细描述中,阐述众多具体细节以实现对本发明概念的透彻理解。然而,应理解,本领域技术人员可以在没有这些具体细节的情况下实践本发明概念。在其它情况下,未详细描述众所周知的方法、过程、组件、电路和网络,以便不会不必要地混淆实施例的各方面。

[0084] 应理解,虽然术语第一、第二等可在本文中用以描述各种元件,但这些元件不应受这些术语限制。这些术语仅用于将一个元件与另一元件区分开来。举例来说,在不脱离本发明概念的范围内,第一模块可以被称为第二模块,并且类似地,第二模块可以被称为第一模块。

[0085] 本文中用于本发明概念的描述中的术语是仅出于描述具体实施例的目的且并不希望限制本发明概念。如本发明概念和所附权利要求书的描述中所使用,单数形式“一(a/an)”和“所述”意在也包含复数形式,除非上下文另外明确指示。还将理解,如本文中所使用的术语“和/或”指代且涵盖相关联的所列项目中的一或多者的任何及所有可能的组合。将进一步理解,当用于本说明书中时,术语“包括(comprises)”和/或“包括(comprising)”指定存在所陈述的特征、整数、步骤、操作、元件和/或组件,但不排除存在或添加一个或大于一个其它特征、整数、步骤、操作、元件、组件和/或其群组。附图中的组件和特征未必是按比

例绘制的。

[0086] 全快闪存储系统由于固态驱动器 (SSD) 价格持续降低且SSD容量持续增大的事实将用于近期的企业数据中心。因此,硬盘驱动器 (HDD) 的优势减弱,将被具有较大电容量的低端SSD,例如多层单元 (MLC) SSD替换。因此,有成本效益的设计将引入多层SSD,其中每一层可能具有独特的特殊性,例如高速、大容量、擅长读取输入/输出 (I/O) 命令、长寿命等。

[0087] 在另一方面,典型的企业数据中心使用范例使用管理程序-虚拟机 (VM) 结构。VM的I/O工作负载模式可以不同,且其升级到更高层服务所发现的性能改进并不相同。由此推动,称作“自动分层 (自动分层)”的整体解决方案可以操控全快闪多层数据中心中的VM配置和迁移以最佳的利用存储资源、优化性能,以及减小迁移开销。自动分层也可以考虑历史性能和预测性能两种因素。

[0088] VM运行客户机操作系统 (OS),例如Linux和Windows,以及应用程序。VM彼此隔离。云服务供应商向用户“出租”其VM。VM基于用户应用可以具有不同的工作负载特征,且因此其对存储装置速度可以具有不同程度的灵敏度。

[0089] VM管理程序软件,例如Xen、KVM和VMware,可以托管单个服务器内的多个虚拟机 (VM)。VM管理程序软件负责调度、资源管理、系统软件应用编程接口 (API) 以及硬件虚拟化。本发明概念的实施例可以包含装设在VM管理程序层级处的自动分层后台程序,所述后台程序的工作与自动分层I/O滤波器和自动分层控制器连通。

[0090] 所有VM服务器可以连接到集中式存储池以共享后端全快闪多层SSD磁盘,所述磁盘可以包含例如非易失性存储器Express (NVMe) SSD、3D XPoint非易失性存储器 (NVM) SSD,和MLC/三阶存储单元 (TLC) /四阶存储单元 (QLC) SSD。每一个层可以具有不同的特殊性,例如快速、较大存储容量等。

[0091] 自动分层可以包含三个组件:自动分层I/O滤波器、自动分层后台程序和自动分层控制器。

[0092] 自动分层I/O滤波器驱动器可以附接到将管理的每一个VM (即,虚拟机磁盘 (VMDK) 文件)。自动分层I/O滤波器驱动器可以收集I/O相关的统计且对每一VMDK运行特定的延迟测试。可以一些样本周期速率收集数据且结果可以传输到负责收集来自所有VMDK的所有数据的主机系统上的自动分层后台程序。

[0093] 自动分层后台程序可以在所有物理机器的VM管理程序上运行。自动分层后台程序跟踪VM服务器的工作负载变化 (I/O存取模式变化),且从自动分层I/O滤波器收集增加延迟测试的结果以传送到自动分层控制器。

[0094] 自动分层控制器可以在专用服务器或存储系统中的嵌入式系统上运行。若VM迁移到另一层以及其对应的迁移开销,则自动分层控制器可以作出决策以基于预测的VM性能触发迁移。

[0095] 从每一个VM的视角来看,某一VM从一个层迁移到另一层的原因是VM可以在迁移之后更好的运行 (较小的平均I/O延迟、每秒更高的I/O操作 (IOPS)、更高的吞吐量等)。但同时,存在相应的迁移成本,这是因为将VM从一个层迁移到另一层相对昂贵且不可忽略。

[0096] 从全局优化的视角来看,由于资源限制和其相应的服务水平协议 (SLA),难以满足所有VM要求同时迁移到其最喜欢层。全局优化应考虑VM随时间的变化以及先前对迁移效应。举例来说,当前最好地分配解决方案由于在运行时间期间改变VM行为可以导致在未来

不良的情况。因此,全局优化应考虑潜在的益处和代偿、迁移开销、每一层上的VM的历史和预测性能、SLA(服务水平协议)等。基于这种信息,自动分层可以试图使用方程式1到方程式6来实现多层全快闪存储系统的最优解决方案。

[0097] 表1.符号。

符号	含义
v	VM 标识, $v \in [1, v_{max}]$, 其中 v_{max} 是最后一个 VM。
t	层标识, $t \in [1, t_{max}]$, 其中 t_{max} 是最后一个层。
$t_{v,\tau}$	VM v 在时期 τ 期间的托管层。
k	不同的资源类型, $t \in [1, k_{max}]$, 例如 IOPS、吞吐量、存储大小等。
τ	时期标识, 其中 $\tau \in [0, +\infty)$ 。
α_k	资源 k 的权重。
β	迁移成本的权重。
$r(v, t_{v,\tau}, \tau, k)$	在层 $t_{v,\tau}$ 上运行的 VM v 的 k 资源使用的预测类型。
$m(v, t_{v,\tau-1}, t_{v,\tau})$	VM v 在时期 τ 期间从层 $t_{v,\tau-1}$ 到层 $t_{v,\tau}$ 的迁移成本。
k_{stg}	存储大小。
$Pr(v, t, \tau), Pw(v, t, \tau)$	在时期 τ 期间 VM v 在层 t 上的读取或写入命令的吞吐量。
$Pr(\Lambda, t, \tau),$	在时期 τ 期间层 t 的所有剩余可用读取或写入吞吐量。

$Pw(\Lambda, t, \tau)$	
$W_{v,\tau}$	VM v 的权重
Γ_k	可以使用的每种类型资源的上限 (以百分比计)。
$R(t_{v,\tau}, k)$	在层 $t_{v,\tau}$ 上的第 k 种类型资源的总容量。

[0100] 最小化:

$$[0101] \quad \sum_{\forall v, \forall t, \forall \tau} w_{v,\tau} \cdot [\sum_{\forall k} \alpha_k \cdot r(v, t_{v,\tau}, \tau, k) - \beta \cdot m(v, t_{v,\tau-1}, t_{v,\tau})] \quad (1)$$

[0102] 进行:

$r(v, t_{v,\tau}, \tau, k) = r_{Prd}(v, t_{v,\tau-1}, t_{v,\tau}, \tau - 1, k) \geq 0,$	$\forall v, \forall t, \forall \tau$	(2)
$r(v, t_{v,\tau_1}, \tau_1, k_s) \equiv r(v, t_{v,\tau_2}, \tau_2, k_s) \geq 0,$	$\forall \tau_1, \tau_2 \in [0, +\infty)$	(3)
$m(v, t_{v,\tau-1}, t_{v,\tau}) = \frac{r(v, t_{v,\tau-1}, \tau - 1, k_s)}{\min(Pr(\Lambda, t_{v,\tau-1}, \tau - 1) + Pr(v, t_{v,\tau-1}, \tau - 1), Pw(\Lambda, t_{v,\tau-1}, \tau - 1))}$		(4)
$t_{v,\tau} \neq \emptyset,$	$\forall v, \forall \tau$	(5)
$\sum_{\forall v} r(v, t_{v,\tau}, \tau, k) \leq \Gamma_k \cdot R(t_{v,\tau}, k),$	$\forall t, \forall k, \forall \tau$	(6)

[0104] 方程式1表示目标函数。主要想法是使整个性能增益减迁移成本:“性能增益-性能损失”最大化。内部“求和”运算符进行每一个VM的所有资源类型(例如IOPS、吞吐量和存储大小等)的使用率的加权和,假设将 v 从层 $t_{\tau-1}$ 迁移到层 t_τ 。注意这里的迁移并不将VM从一个

主机服务器迁移到另一个主机服务器:仅后端VMDK文件从一个SSD层迁移到另一个SSD层。因此,不需要考虑非磁盘I/O相关的资源,例如CPU、存储器等。外部“求和”运算符进一步迭代所有可能的迁移情况。注意权重参数 $w_{v,\tau}$ 可以反映每一个VM的SLA。

[0105] 方程式2表示可以用于预测性能增益的预测模型函数。方程式3确存储大小(即,VMDK大小)在每一个迁移之前和之后将不会改变。方程式4估计时间迁移开销。详细地说,时间迁移开销是待迁移的VM大小除以通过在对对应层上读取和写入数据所施加的瓶颈。详细地说, $Pr(\Lambda, t_{v,\tau-1}, \tau-1)$ 表示可用的剩余速度,且由于VM在迁移期间可以暂停,所以这个VM所使用的吞吐量也可用(即, $Pr(\Lambda, t_{v,\tau-1}, \tau-1)$ 可以加回)。 $Pw(\Lambda, t_{v,\tau-1}, \tau-1)$ 表示呈写入速度的迁移。方程式5反映需要保证VM被一个且仅一个磁盘层代管。最后,方程式6表示资源限制,其中 Γ_k 是可以使用的每种资源类型的预设上限(以百分比计)。

[0106] 由于以下挑战,很难在运行时间期间对所有时期进行全局优化:

[0107] 1) 系统几乎没有关于未来工作负载模式的信息,所以很难在运行时间期间对所有 τ 时段进行全局优化。

[0108] 2) 当执行全局优化时,对于每一个时期的决策并非无关,这意味着传统的次优类动力编程技术可能无法用于通过优化每一个时期的目标函数值“性能增益-性能损失”来解决问题。

[0109] 3) 根据性能预测模型的复杂度(所述性能预测模型预测在不迁移的情况下在其它层上的VM性能),优化问题可以容易地变成NP难题。

[0110] 因此,为了在可接受的时间内解决这个问题,一些限制需要放宽。

[0111] 代替试图对所有时间执行全局优化,可以寻求仅对每一个时期优化的解决方案(即,运行时间贪婪算法)。

[0112] 随后的问题类似于加权背包问题。另外,由于每一个层的性能“专业”(例如高IOPS、高吞吐量、大空间、较小的写入放大因子(WAF)函数、大量过度配置和较大编程/擦除(P/E)周期等)事先已知,所以可以计算关于每一个VM的以专业资源维度(例如层1的IOPS和吞吐量)为重点的估计性能的“得分”,并且基于性能专业得分排序作出迁移决策,以及对迁移开销的相应估计。

[0113] 可以安排两个不同的时期:监测期,表示评估性能的频率,且回归性能估计模型,其可设定为任何所需的间隔:例如15分钟;和迁移期,表示可以执行从一个层到另一个层的VM迁移的频率,其可设定为任何所需的间隔:例如1天或1周。监测期和迁移期不一定固定:其可以使用能够适应性地调节其在运行时间期间的大小的函数来确定。迁移期也不需要是监测期的整数倍数。其窗口大小越小,监测并测量存储系统且迁移VM就越频繁。系统管理员可以通过在使用之前进行灵敏度分析来平衡精确度与迁移成本之间的折衷。

[0114] 为了在不迁移的情况下估计VM在其它层上的性能,可以人工地将额外延迟增加到每一个VM的I/O命令且通过调用自动分层I/O滤波器API(例如,用于I/O过滤的VMware vSphere API)测量对总I/O延迟产生的效果来“仿真”VM在另一层上的性能。

[0115] 可以使用线形曲线模拟VM在不同层中的性能改变。平均I/O延迟对比所增加的延迟的曲线可以回归成线形方程式。线条的斜率可能取决于当时正在运行的应用程序是同步或异步。同步应用程序和异步应用程序在其如何操作上不同,其反映了线条的斜率。具体地说,当同步应用程序传送I/O命令时,同步应用程序等待I/O命令完成,这会在通信正在进行

时阻断应用程序的进度,并使系统资源闲置。相比之下,异步应用程序可以传送I/O命令并且可以在传输已结束之前执行其它处理。这些事实意味着异步应用程序倾向于受I/O命令延迟的较小影响,但改进出现的代价是:异步应用程序并不知道响应将何时传输,且必须能够在任何时间接收所述响应。

[0116] VM对层速度的灵敏度不同。具有同步应用程序的VM比具有异步应用程序的VM对层速度变化更具有灵敏性。主要原因是异步应用程序通常具有不太频繁的I/O命令待刷新,且对I/O时序具有更小的依赖。

[0117] 由这些观测结果推动,“层速度灵敏度校正会话”(TSSCS)可以用于通过在不真正迁移的情况下迁移VM到更高(或更低)速度层来预测所述VM可以增益(或损失)多少。详细地说,TSSCS具有以下特性:

[0118] 1) 轻量级:在自动分层I/O过滤器上运行,TSSCS将一些延迟增加到每一个VM的I/O命令中,且花费极短的时间以便不影响当前的托管工作负载。

[0119] 2) 多个样本/延迟/TSSCS:为改进在每一个层下的每一个VM性能的仿真精确度,可以平均在每一个层的相同增加延迟下所获得的结果。

[0120] 3) 多个延迟/TSSCS:为优化回归,可以增加多个不同的延迟。此外,所增加的延迟并不必须受限于当前存储池中的可用层。实际上,额外的延迟数据点可以用于提高回归精确度。

[0121] 4) 在运行时间期间的多个TSSCS:TSSCS可以被周期性触发以更新曲线和回归函数。

[0122] 基于相应的性能,性能增益与层速度之间的相关性可以通过线性回归来确定,所述相关性反映了每一个VM对磁盘速度的灵敏度提高或降低。曲线有可能回归成多项式或非线性函数,但改变回归并不影响自动分层如何操作。给定当时的每一个VM的增加延迟曲线,可以计算每一个VM在每一个层的IOPS、吞吐量、存储大小等。注意到虽然IOPS、吞吐量和存储大小为焦点,但其它类型的资源也可考虑。

[0123] 表2概述这个分段中所使用的符号,且表3进一步示出使用在三个不同层下的三个VM的实例的估计平均I/O延迟的结果。

[0124] 表2.符号。

[0125]	符号	含义
	L_t	层 t 的初始平均 I/O 延迟 (无增加延迟)
	b_v, m_v	用于 VM x 的经回归层速度敏感度相关性模型的两个参数, 即 $y = mx + b$, 其中 y 为平均 I/O 延迟且 x 为增加的延迟。
	s_v	VM v 的平均 I/O 大小。
	$S_v, VM[v].size$	VM v 的存储大小。
	$maxIOPS[t], maxMBPS[t], maxSize[t]$	层 t 的预设可用资源上限, 其中我们设定所述上限为每一种类型资源的总容量的 90%。

[0126] 表3.在不迁移的情况下估计平均I/O延迟的实例

[0127]		层1	层2	层3
	V1平均I/O延迟 (μs)	b_1	$m_1 (L_2 - L_1) + b_1$	$m_1 (L_3 - L_1) + b_1$

V2平均I/O延迟 (μs)	$m_2(L_1 - L_2) + b_2$	b_2	$m_2(L_3 - L_2) + b_2$
V3平均I/O延迟 (μs)	$m_3(L_1 - L_3) + b_3$	$m_3(L_2 - L_3) + b_3$	b_3

[0128] 层速度灵敏度校正会话 (TSSCS) 可以迭代存储系统中的每一个VM和每一个层。所述层速度灵敏度校正会话可以针对每一个增加延迟计算样本的I/O延迟的系数变量 (即, CV)。随后可以针对每一个增加延迟计算I/O延迟的平均值。随后可以求所有增加延迟的CV值的平均值, 且基于结果, 自动分层可以界定用于产生每一个VM的线性回归的增加延迟的置信度, 以及因此随后自身线性回归中的置信度。CV越低, 这种VM具有的置信度就越高。若CV大于1, 则置信度可以设定为可以根据系统管理员的要求来调节的下限: 例如0.05。最后, TSSCS可以调用回归函数来计算相应的m值和b值。注意到虽然本发明概念的实施例可以使用在返回元组中具有仅两个参数的线性回归模型, 但系统管理员可以基于其选择的回归模型添加更多项目到返回元组中。

[0129] 自动分层具有三种二维矩阵来记录三种类型的资源, 例如IOPS (I/O命令/秒)、吞吐量 (数兆字节/秒或MBPS) 和存储大小 (字节)。可以视需要添加用于存储系统中的任何其它类型的资源的额外矩阵。与IOPS和吞吐量相比, 用于存储大小的矩阵相对地简单通过调用管理程序API测量每一个VM占用的存储大小来得到。

[0130] 自动分层可以使用基于经回归M值和B值的线形估计来估计IOPS。在得到用于计算不同层上的每一个VM的相应IOPS和吞吐量的估计平均I/O延迟结果后, 可以使用以下方程式来计算IOPS和吞吐量:

$$[0131] \quad IOPS = \frac{\text{平均IO大小}(v) \times 10^6}{\text{平均IO延迟}(v, t)} \quad (7)$$

[0132] 吞吐量 = 平均IO大小 $\times$ IOPS (8)

[0133] 由于不同类型的资源具有不同的单位, 所以难以将其一起放置在单个多维矩阵中。因此, 每一种资源可以转换成每一个层的总的可用资源容量的百分比。这种归一化的实例示出在表4中。

[0134] 表4.3个VM在层1上运行时的归一化性能矩阵的实例

	IOPS	吞吐量	存储
[0135] VM1	$\frac{10^6}{b_1 \times \max IOPS[t]}$	$\frac{s_1 \times 10^6}{b_1 \times \max MBPS[t]}$	$\frac{S_1}{\max Size[t]}$
VM2	$\frac{10^6}{[m_2(L_1 - L_2) + b_2] \times \max IOPS[t]}$	$\frac{S_2 \times 10^6}{[m_2(L_1 - L_2) + b_2] \times \max MBPS[t]}$	$\frac{S_2}{\max Size[t]}$
[0136] VM3	$\frac{10^6}{[m_3(L_1 - L_3) + b_3] \times \max IOPS[t]}$	$\frac{s_3 \times 10^6}{[m_3(L_1 - L_3) + b_3] \times \max MBPS[t]}$	$\frac{S_3}{\max Size[t]}$

[0137] 计算出的得分应反映以下事实:

[0138] 1) 层和VM两者的特征: 如果在每一个层上运行, 则得分应反映每一个层的特殊性和每一个VM的工作负载特征。因此, 单独地计算每一个层上的每一个VM的得分。

[0139] 2) SLA权重: VM由于其具有不同的SLA权重而不相同, 如方程式1中所示。

[0140] 3) 估计的置信度: 在性能矩阵中计算的系数变量可以用于反映估计的置信度。

[0141] 4) 历史成本和迁移成本: 卷积老化因子可以用于将历史得分包含到得分计算值中。迁移成本也可以被估计且包含在得分计算值中。

[0142] 基于用户配置或默认配置, 由每一个层提供的资源为事先已知的。自动分层可以对给定层的多种资源进行优化。为反映特殊性, 可以使用层特殊性矩阵。实例层特殊性矩阵示出于表5中。在表5中, 针对特殊性配置使用所有间隔尺寸或不使用间隔尺寸。具有1的存储单元意味当前层在相应类型的资源中具有特殊性。否则, 所述层不针对所述类型的资源进行优化。但在本发明概念的其它实施例中, 这个矩阵中的值可以不一定为0和1。如果系统管理员想要更细粒度的间隔尺寸来控制特殊性, 则其可以设定任何权值代替0和1。系统管理员也可以通过添加更多类型的资源 (即, 更多行) 来延伸这个矩阵。

[0143] 表5. 层特殊性矩阵的实例。

[0144]

	IOPS	吞吐量 (MBPS)	存储大小 (字节)
层1	1	1	0
层2	0	1	0
层3	0	0	1

[0145] 在通过默认最昂贵到最不昂贵层的次序来分选所述层时是有帮助的, 且这种次序被视为在迁移决策制定时段期间的优先次序。注意到系统管理员也可以基于其对每一个层的偏好来改变这种次序。

[0146] 下一个问题是“如何在反映每一个层的特殊性时反映每一个VM在每一个层上的性能?”。由于术语“性能”考虑多个维度或类型的资源 (例如IOPS、吞吐量和存储大小), 所以加权总和用于反映每一个层的一种或多种特殊性。可以使用“正交匹配”, 其是对“特殊性”矩阵和VM容量利用率矩阵的每一层每一VM的相乘操作, 如方程式9中所示:

[0147]
$$\begin{bmatrix} \text{VM容量率矩阵}[t][v].\text{IOPS} \\ \text{VM容量率矩阵}[t][v].\text{MBPS} \\ \text{VM容量率矩阵}[t][v].\text{大小} \end{bmatrix} \times [\text{特殊性}[t].\text{IOPS}, \text{特殊性}[t].\text{MBPS}, \text{特殊性}[t].\text{大小}] \quad (9)$$

[0148] 另外, 自动分层还允许系统管理员为每种类型的资源配置不同的权值。

[0149] 计算层1的容量得分的实例示出于表6中。第一行 (“容量率”) 中的 (IOPS%, MBPS%, 大小%) 元组是对这个层上的每种类型的资源的估计VM利用率。在这个实例中, 这些值中无一者大于1, 所以全部三个VM都通过限制测试。最后, 假设特殊性矩阵如表5中所示。

[0150] 表6. 3个VM在层1上运行时的容量率、容量限制和容量得分值的实例。

[0151]

	VM1	VM2	VM3
容量率	(0.5, 0.6, 0.3)	(0.2, 0.3, 0.3)	(0.1, 0.5, 0.4)
容量限制	通过	通过	通过
当前容量得分	$(0.5+0.6)/2=0.55$	$(0.2+0.3)/2=0.25$	$(0.1+0.5)/2=0.30$

[0152] 最终得分是历史得分、当前得分和迁移成本的卷积总和, 如方程式10中所示:

[0153]
$$\text{总得分} = \text{老化因子} \times \text{历史得分} + \text{当前容量得分} - \text{迁移权重} \times \text{迁移成本} \quad (10)$$

[0154] 这个最终得分具有三种组分:

[0155] 1) 历史得分: 由于仅基于最新时期作出决策, 为了避免一些VM在两个层之间频繁

地迁移且切换,自动分层考虑历史得分,以预设老化因子来衰减旧得分。最初,当前未分配到特定层的任一VM将接收零得分。

[0156] 2) 当前容量得分:当前容量得分参考方程式(9)和表6如上文所描述计算。又,当前未分配到特定层的任一VM将接收零得分。

[0157] 3) 迁移成本:自动分层通过选择迁出和迁入可用吞吐量的较小值来计算迁移速度瓶颈。自动分层随着迁移时间成本传回VM大小除以迁移速度的结果。

[0158] 为计算得分,针对每一个层,自动分层可以迭代其托管VM并检查当前VM是否超过当前层t的任何类型资源的上限。若VM超过任一资源容量上限,则相应的VM容量利用率可以设定为0,且层t上的这个VMv的得分设定为-1。否则,将资源使用率转换成容量百分比。随后如上文计算总得分。注意到自动分层亦允许系统管理员通过调节资源的权重值来控制每种类型资源的权重。

[0159] 自动分层在“长期视图”中进行迁移。也就是说,自动分层遵循相对较低的频率来迁移,这是因为迁移操作相对昂贵。

[0160] 自动分层可以检查迁移操作是否需要且在必要时可以触发相应的迁移。自动分层可以迭代每一个层,优选地以最昂贵到最便宜层的次序。注意到系统管理员可以基于其对每一个层待偏好来改变这种次序。随后自动分层按降序分选层得分。对于这个分选清单中的每一个VM,若VM在以下情况下,则自动分层可以跳过所述VM:

[0161] 1) 当前迭代VM在当时已经被分配(迁移),或

[0162] 2) 由于这个VM超过了这个层的容量限制,所以所述层上的所述VM没有得分,或

[0163] 3) 在分配期间层t上没有充足的资源容量用于这个VMv。

[0164] 否则,自动分层可以将VM从最多得分到最少得分逐一地分配到这个层。注意到若VM已经在其目标层上,则这种分配不会迁移所述VM。一旦VM被分配另一层后,则所述VM可以标记用于迁移,其包括从源存储层复制数据(VMDK文件)到目标存储层。

[0165] 常规数据中心使用HDD存储数据。HDD在大多数意义上通常与彼此相当类似,主要不同在于其每千兆字节的成本,其读取/写入速度和其整体容量。正如可预期的那样,容量越大或读取/写入速度越快,HDD就越昂贵。但即使考虑到其差异,HDD也具有平均\$0.25/千兆字节的成本,且具有平均10毫秒(ms)的读取/写入速度。但甚至10ms在电脑操作中也可以是较长的延迟,通常使得一些数据被局部缓存以提高存取时间。

[0166] 在另一方面,SSD可以提供不同的专业化。一些SSD可以具有较低的读取/写入延迟,其它SSD可以具有高带宽,又其它SSD可以具有大容量,且又其它SSD可以具有较小的写入放大因子。在本文中,术语“资源”用于描述单个存储装置可提供到更大程度或更小程度的这些专业化。其它SSD仍可以提供这些资源的组合。这些额外专业化为应用程序提供可以利用这些资源的益处,但使确定哪个SSD对于特定应用是最好的方法复杂化。

[0167] 与HDD相比,SSD通常更昂贵,其中视格式而定,每千兆字节的价格在\$0.28到\$4.68的范围内变化。但SSD通常还极具响应性:根据将使用的SSD格式,读取/写入时间可以在1-2ms到微秒(μ s)或甚至纳秒(ns)的范围内变化。

[0168] 虽然SSD通过HDD提供的附加资源是如何最好地将应用程序分配给SSD的重要因素,但存取时间是对应用程序可如何分配给SSD的影响最大。常规缓存算法假设存储层之间的性能得到相当大的改进:通常根据本地缓存层相对于HDD存储,预期性能得到10倍改进。

此外,鉴于SSD比HDD存储数据每千兆字节更昂贵,缓存数据将导致数据以一定的成本存储两次。因此,在使用除HDD以外的存储装置时,常规的分层和缓存算法产生次优结果。

[0169] 图1示出根据本发明概念的实施例的可以托管虚拟机并且使用经装备以执行自动分层的存储池的连网机器。在图1中,示出系统105。系统105可以包含虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3。虽然图1示出了三种服务器,但本发明概念的实施例可以支持任何数量的虚拟机服务器。

[0170] 虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3中的每一个可以分别包含处理器115-1、处理器115-2和处理器115-3,且分别包含存储器120-1、存储器120-2和存储器120-3。在其它使用中,处理器115-1、处理器115-2和处理器115-3可以执行应用程序,例如虚拟机在虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3上运行所使用的应用程序,且所述应用程序可以存储在存储器120-1、存储器120-2和存储器120-3中。处理器115-1、处理器115-2和处理器115-3可以是任何品种的处理器的处理器:例如Intel Xeon、Celeron、Itanium或Atom处理器、AMD Opteron处理器、ARM处理器等。虽然图1示出了虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3中的每一个中的单一处理器,但虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3可以各自包含任何数量的处理器,其中的每一个可以是单核处理器或多核处理器,且可以任何所需的组合混合。存储器120-1、存储器120-2和存储器120-3各自可以是任何品种的存储器,例如快闪存储器、动态随机存储器(DRAM)、静态随机存取存储器(SRAM)、永久性随机存取存储器、铁电随机存取存储器(FRAM),或非易失性随机存取存储器(NVRAM),例如磁阻随机存取存储器(MRAM)等。存储器120-1、存储器120-2和存储器120-3各自可以是不同存储器类型的任何所需组合。存储器120-1、存储器120-2和存储器120-3可以由存储器控制器(在图1中未示出)控制,所述存储器控制器也是虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3的一部分。

[0171] 虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3各自还可以包含本地存储装置。举例来说,虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3可以分别包含缓存125-1、缓存125-2和缓存125-3。例如125-1、缓存125-2和缓存125-3的本地存储装置可以采用任何所需形式,例如硬盘驱动器(HDD)或固态驱动器(SSD),且可以使用任何所需的连接器,例如外围组件互连(PCI)、PCI Express(PCIe)、串行AT配件(SATA)、非易失性存储器Express(NVMe)、网状架构的NVMe(NVMeoF)、M.2,或任何其它所需连接器。在本发明概念的一些实施例中,虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3可以忽略缓存125-1、缓存125-2和缓存125-3,仅依赖数据中心中的存储。在包含缓存125-1、缓存125-2和缓存125-3的本发明概念的实施例中,本发明概念的实施例可以与公开于2018年2月7日提交的美国专利申请案第15/891350号中的本发明概念的实施例组合,所述美国专利申请案要求2017年11月29日提交的美国专利申请案第62/592,355号的优先权,其两者均出于所有目的以引用的方式并入本文中。

[0172] 系统105还可以包含数据中心服务器130。类似于虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3,数据中心服务器130可以包含处理器115-4和存储器120-4。但数据中心服务器130还可以包含存储池135、接收器140和传输器145。存储池135可以是经组织成存储层的存储装置池。在本发明概念的一些实施例中,存储池135可以包含仅SSD或其它快闪存储器装置;在本发明概念的其它实施例中,存储池135可以包含使用其它存储格

式的存储装置。接收器140可以接收输入/输出(I/O)命令和来自虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3的其它数据,且传输器145可以将数据传输到虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3。

[0173] 虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3,与数据中心服务器130经示出已经由网络150连接。使用网络150,虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3可以将I/O命令传送到数据中心服务器130,且可以自其接收响应。虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3与数据中心服务器130可以全部是局域网(Local Area Network;LAN)的一部分,或可以包含其它网络类型,例如广域网(Wide Area Network;WAN)或例如因特网的全球网络。网络150还可以是多个这类网络的组合,呈任何所需组合的形式。

[0174] 尽管图1描绘虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3以及数据中心服务器130作为服务器(且可以各自为独立式或机架式服务器),但本发明概念的实施例可以包含各自具有任何所需类型而不受限的虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3以及数据中心服务器130。举例来说,图1所示的服务器可以各自被台式或膝上型计算机或可以得益于本发明概念的实施例的任何其它装置替换。虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3以及数据中心服务器130各自还可包含特定得便携式计算装置、平板计算机、智能手机和其它计算装置。出于描述的目的,虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3被描述为托管虚拟机,但本发明概念的实施例可以包含在虚拟机服务器110-1、虚拟机服务器110-2和虚拟机服务器110-3上运行的不是虚拟机的机器。

[0175] 图2示出图1的服务器110-1、服务器110-2、服务器110-3和服务器130的额外细节。在图2中,通常地,服务器110(其可以表示图1的服务器110-1、服务器110-2或服务器110-3中的任一个)和服务器130包含一个或多个处理器115,所述处理器可以包含存储器控制器205和时钟210,所述时钟可以用于协调服务器110和服务器130的组件的操作。处理器115还可以耦合到存储器120,举例来说,所述存储器可以包含随机存取存储器(random access memory,RAM)、只读存储器(read-only memory,ROM)或其它状态保持介质。在服务器130中,处理器115还可以耦合到存储装置135(服务器110还可以包含局部附接的存储装置,但这类装置不是所需的)。在服务器110和服务器130中,处理器115还可以耦合到网络连接器215,所述网络连接器可以是例如以太网连接器或无线连接器。处理器115还可以连接到总线220,在其它组件中,用户接口225和可以使用输入/输出引擎230管理的输入/输出接口端口可附接到所述总线。

[0176] 图3示出根据本发明概念的实施例的图1的虚拟机服务器110与图1的存储池135在自动分层的情况下进行交互的附加细节。在图3中,虚拟机服务器110示出托管三个虚拟机305-1、虚拟机305-2和虚拟机305-3,但本发明概念的实施例可以包含由虚拟机服务器110托管的任何数量的虚拟机。虽然仅虚拟机305-3在下文更详细地描述,但本发明概念的实施例包含其它虚拟机,例如类似操作的虚拟机305-1和虚拟机305-2。虚拟机305-3示出为包含在操作系统315上运行的应用程序310。管理程序320可以管理在虚拟机服务器110上运行的操作虚拟机305-1、操作虚拟机305-2和操作虚拟机305-3。

[0177] I/O滤波器325可以充当传递滤波器,接收来自应用程序310的命令并将其传送到

管理程序320用于执行。I/O滤波器325可以是任何所需的I/O滤波器,例如用于I/O过滤的VMware vSphere API。但除其正常操作以外,I/O滤波器325可以将延迟增加到一些来自虚拟机305-3的I/O命令中,所述虚拟机可以提供关于VM在由其它存储层支持时如何估计其性能的信息。I/O滤波器325的操作参考以下的图7进一步论述。

[0178] 自动分层后台程序330可以接收来自I/O滤波器325的性能数据并将其传送到自动分层控制器335。自动分层后台程序330可以是管理程序320的一部分,或其可以是独立组件。另外,虽然自动分层后台程序330可仅是简单的传递装置,接收来自I/O滤波器325的数据并将其传送到自动分层控制器335,但自动分层后台程序330可以包含稍后归因于自动分层控制器335的功能中的一些,例如对关于虚拟机305-1、虚拟机305-2和虚拟机305-3的性能数据执行线性回归的操作。而在本发明概念的一些实施例中,自动分层后台程序330可以完全被省略,其中I/O滤波器325直接与自动分层控制器335连通而不是准许自动分层后台程序330收集并传送性能数据。

[0179] 自动分层控制器335负责取得从自动分层后台程序330接收的性能数据并使用所述数据以估计虚拟机305-3将在每一个存储层上如何执行,之后自动分层控制器335可以按需要将数据从虚拟机305-3迁移到另一存储层。由自动分层控制器335管理的存储池135可以包含任何数量的存储层:图3示出三个层340-1、层340-2和层340-3。在每一个层内,可存在任何数量的存储装置345。

[0180] 如上文所提及,每一个存储层可以提供利用由所述存储层提供的特定资源的专业化。举例来说,存储层340-1可以提供快速的读取/写入时间,而存储层340-3可以提供较小的写入放大因子。由于存储层340-1、存储层340-2和存储层340-3提供专业化,所以所述存储层内的存储装置345应全部提供专业化。因此,在本发明概念的一些实施例中,给定存储层内的存储装置345相对于一个或多个特殊性而言是相对或完全均质的:在一些情况下,在给定存储层内的存储装置345和所有存储装置345应是相同模式的存储装置。

[0181] 在本发明概念的一些实施例中,自动分层控制器335实施为图1的数据中心服务器130的一部分。在本发明概念的其它实施例中,自动分层控制器可以实施为可能在虚拟机服务器110-1虚拟机服务器110-2和虚拟机服务器110-3中的一个上运行的虚拟机。

[0182] 图4示出由不同存储装置提供的不同可用资源。在图4中,示出资源405。资源405可以包含吞吐量410,所述吞吐量表示可以在给定时间量例如一秒中执行的I/O命令的数量(通常描述为I/O操作/秒,或IOPS);带宽415,所述带宽反映在给定时间单位内可以传输多少数据(通常以数兆字节/秒,或MBPS描述);容量420,所述容量反映可以存储多少数据;以及写入放大因子425,所述写入放大因子反映由于垃圾收集操作而写入多少额外数据。资源405可以容易扩展以包含由图1的存储池135提供的任何其它所需资源。

[0183] 图5示出将从一个存储装置迁移到另一个存储装置的图3的虚拟机的文件。在图5中,虚拟机文件505存储虚拟机的数据。虚拟机文件505当前驻存在存储层340-3上,而第二虚拟机的虚拟机文件510目前驻存在存储层340-2上。如箭头515所示,虚拟机文件505可以迁移到存储层340-2或存储层340-1,根据何种迁移,图3的自动分层控制器335将会引起图1的系统105整体的最好提高。注意到图3的自动分层控制器335可以决定不迁移虚拟机到另一存储层,如虚拟机文件510在存储层340-2上保持不迁移的事实所示。

[0184] 虽然最好存在一个如何将虚拟机分配到不变(至少,只要虚拟机集群不改变)的存

储层的最优解决方案,但实际上虚拟机可能经历工作负载改变,且可以与其它VM(其需求也可以随时间改变)竞争资源。工作负载和竞争的这些变化可以影响哪个存储层对于特定虚拟机来说是“最好的”。举例来说,在一个时间点,一个虚拟机可以非常忙且需要来自图1的存储池135的快速响应时间,但在另一时间点,同一虚拟机可以暂停工作且不需要高速存储层。因此,可能不存在一直为所有虚拟机最好工作的单一全局优化。

[0185] 因此,代替寻求单一的全局优化,优选的是寻求一种在一段时间内工作的优化,之后可能优选的是不同的优化。举例来说,图6A到图6B示出在不同时间点虚拟机到存储层的最优分配。举例来说,在图6A中,在时间T0处,虚拟机文件505最优在存储层340-3中且虚拟机文件510最优在存储层340-1中。相比之下,在图6B中,在时间T1处,虚拟机文件505和虚拟机文件510两个最优在存储层340-1中。

[0186] 图7示出增加延迟并且确定图3的虚拟机的性能数据的图3的I/O滤波器325。在图7中,I/O滤波器325可以接收来自图3的虚拟机305-3的应用程序310的I/O命令705。I/O滤波器325随后可以将延迟710增加到I/O命令705中。这种增加延迟可以模仿更慢执行存储的层,所述更慢执行存储的层针对虚拟机产生不同的平均I/O延迟。

[0187] 使用所增加的延迟710,I/O滤波器325可以提供I/O命令705到数据中心服务器130。数据中心服务器130随后可以处理I/O命令705并返回响应715。I/O滤波器325可以返回响应715到应用程序310并且可以使用所增加的延迟710测量图3的虚拟机305-3的平均I/O延迟,且产生性能数据720。图3的自动分层控制器335随后可以使用性能数据720来确定是否存在将虚拟机迁移到另一存储层(更快或更慢,这是因为迁移虚拟机到更慢的存储层可以为另一个具有性能补偿改进的虚拟机释放空间)的优势。

[0188] 虽然图7示出使用所增加的延迟710以产生性能数据720的I/O滤波器325,但所增加的延迟710具有其自身的开销。视需要,所增加的延迟710可以用轻量路径替换,条件是产生性能数据720。举例来说,I/O滤波器325可以跟踪未完成I/O命令的数量/虚拟机,且使用所述信息以提出在未完成的I/O命令数量仍不断地增高或保持递增时将虚拟机迁移到另一存储层。

[0189] 图8示出图3的自动分层控制器335的细节。在图8中,自动分层控制器335可以包含线性回归模块805、资源矩阵生成器810、度量转换器815、专业化信息820、层选择模块825和迁移模块830。线性回归模块805可以采用图7性能数据720并对所述数据执行线性回归,产生估计虚拟机对于给定存储层的延迟的平均I/O延迟的线形方程式。如下文参考图10所描述,给定增加的延迟,可以确定给定虚拟机的平均I/O延迟,并将这些数据点绘制在曲线上。在本发明概念的一些实施例中,这些数据点是近似地线形,且线性回归可以产生近似给定的增加延迟的平均I/O延迟的方程式。随后,已知每一个存储层的延迟,可以针对每一个存储层估计给定虚拟机的平均I/O延迟。

[0190] 资源矩阵生成器810可以采用可以由使每一个虚拟机与每一个存储层和由每一个存储层提供的资源相关联的线性回归产生的估计。举例来说,可以根据虚拟机的平均I/O大小和虚拟机的平均I/O延迟来计算吞吐量,且可以根据VM的I/O命令的平均I/O大小和VM发出的IOPS计算带宽。资源矩阵生成器810可以使用度量转换器815将一个度量的性能数据(例如吞吐量)转换成另一度量的性能数据(例如带宽)。

[0191] 专业化信息820可以指示每一个存储层提供什么资源。专业化信息进一步参考下

图11来论述。

[0192] 层选择模块825可以采用专业化信息820和由资源矩阵生成器810产生的资源矩阵,并且选择虚拟机的存储层。随后,如果虚拟机将从一个存储层迁移到另一个存储层,则迁移模块830可以执行所述迁移。

[0193] 图9示出将图7的性能数据720转换成使虚拟机、存储层和所述存储层的资源相关联的矩阵。在图9中,性能数据720表示为在使用给定存储层时每一个虚拟机的平均I/O延迟的矩阵。因此举例来说,估计虚拟机1在使用存储层1时具有800 μ s的平均延迟,在使用存储层2时具有500 μ s的平均延迟,且在使用存储层3时具有300 μ s的平均延迟,其中虚拟机2和虚拟机3的性能数据提供类似的信息。图8中所示出的数值仅为示范性的,且不必反映任何虚拟机在存储层中的真实性能值,这是因为实际性能很大程度上取决于特定虚拟机将运行的应用程序和其它虚拟机从同一存储池存取数据的影响两者。

[0194] 注意到性能数据720中的一些值不是估计值而是实际值。具体地说,每一个虚拟机将其数据存储在图1的存储池135中的一些存储层上。因此,对于实际存储虚拟机的数据的存储层来说,性能数据720中的平均I/O延迟表示所述虚拟机在所述存储层上的实际平均I/O延迟,而不是估计值。

[0195] 给定每一个虚拟机在每一个存储层上的估计平均延迟,可以产生使虚拟机、存储层和资源相关联的多维矩阵。这种多维矩阵在图9中示出为资源矩阵905。由于很难在纸上示出多维矩阵,所以图9将资源矩阵905示出为若干单独的二维矩阵910、二维矩阵915和二维矩阵920,其中每一个二维矩阵存储单个存储层的数据,但这仅出于观测目的。

[0196] 在存储层1的二维矩阵910中,根据性能数据720的平均I/O延迟可以使用公式

$$\frac{10^6}{AverageIOlatency * maxIOPS[t]}$$
转换成吞吐量:也就是说,106(1,000,000)除以所讨论的使用所述存储层的虚拟机的平均I/O延迟与存储层可以支持的最大IOPS(这个值可以由系统管理员设定为小于存储层所提供的真正最大值的一些值:例如,真正最大值的90%)的乘积。可以使用值 10^6 ,这是因为延迟通常以微秒来表达(一微秒= $\frac{1}{1,000,000}$)。因此,除非将

比如12 μ s的延迟转换成0.000012秒,否则平均I/O延迟的精确测量应反映方程式中出现的单位变化。吞吐量可以通过用吞吐量乘以由虚拟机发出的I/O命令的平均写入大小来转换成带宽。并且容量可以根据虚拟机文件的存储大小除以由存储层提供的最大存储来计算(另外,系统管理员可以将这个值设定成小于真正最大值的一些值)。可以使用图8的度量转换器815计算从性能数据720到存储于资源矩阵905中的值的这些转换。

[0197] 图10示出图7的性能数据720的线性回归。在图10中,示出两种不同虚拟机的性能数据。对于一个虚拟机,曲线上的数据点可以通过线性回归1005来粗略估计;对于另一个虚拟机,曲线上的数据点可以通过线性回归1010来粗略估计。虽然图10示出仅两个虚拟机的数据点,但本发明概念第实施例可以支持线性回归可以计算的任何数量的虚拟机。

[0198] 每一个数据点使特定的增加延迟与虚拟机的平均I/O延迟相关联。因此,举例来说,当没有增加的延迟时,第一虚拟机经历3000 μ s的平均I/O延迟,而第二虚拟机经历近似700 μ s的平均I/O延迟。

[0199] 增加的延迟尤其在确定虚拟机是否会使用较慢存储层(即,具有较低性能水平的

存储层)来充分操作时起作用。毕竟,增加延迟的量可以正好是当前存储层与目标存储层之间的差。但使用增加的延迟不可能仿真较快存储层:毕竟,不可能增加负值延迟,且存储层不可能比其实际能够响应更快速地响应。解决方案如线性回归1010所示。使用关于虚拟机如何响应较慢存储层的信息,有可能使用线性回归模型估计更快存储层的性能。在技术上,与“内插法”相比较,估计所提供数据范围之外的值称作“外推法”,所述内插法是对数据范围内的值进行估计。但当使用例如线性回归1010的线性回归时,计算对于内插法和外推法是相同的。因此,使用线性回归1010,有可能估计虚拟机在使用更快存储层时将会如何响应。

[0200] 虽然有可能使用仅对应于存储层之间的实际延迟差的数据点,但更多数据点是可用的,由线性回归模块805提供的估计值有可能更好。因此,可以使用图3的I/O滤波器325增加各种不同的延迟以视需要提供为多个数据点。

[0201] 当然,如果虚拟机的每一个I/O命令使用增加的延迟而减缓,那么虚拟机将不接收经分配存储层的真正益处。因此,所增加的延迟可以仅用于由虚拟机发出的I/O命令子集。如果选择以增加延迟的I/O命令是随机选择的,那么图7的性能数据720应提供有意义的线性回归1005和线性回归1010。已知的统计技术可以用于确定数据点的变量系数(和置信度)。如果数据点中的置信度足够高,那么可以信任线性回归1005和线性回归1010为精确估计值。可以指定任何所需的置信度:例如,90%或95%。参考下图13进一步论述如何选择I/O命令用于增加延迟。

[0202] 注意到线性回归1005和线性回归1010完全非常不同的数据集:线性回归1005包含极其陡峭的斜率,而线性回归1010包含极其浅的斜率。并非所有虚拟机以相同的方式响应其I/O命令的延迟变化:一些虚拟机比其它更敏感。举例来说,产生得到线性回归1005的数据点的虚拟机可以运行同步应用程序,其中延迟的较小变化可以具有对应用程序性能的显著影响。在另一方面,产生得到线性回归1010的数据点的虚拟机可以运行异步应用程序,其中延迟的变化具有较小影响。或者,线性回归1005可以用于需要高吞吐量的虚拟机,而线性回归1010可以用于需要低吞吐量的虚拟机。以类似的方式,虚拟机的灵敏度可以影响其线性回归:更敏感的虚拟机将具有更陡的斜率,且不太敏感的虚拟机将具有更浅的斜率。针对灵敏度的类似概念是虚拟机是紧密耦合或不是紧密耦合(这里的“紧密耦合”是指虚拟机对存储层的响应时间有多敏感或依赖,类似于同步应用程序与异步应用程序之间的比较):虚拟机越紧密耦合,其线性回归的斜率将越可能更陡。一般来说,具有陡峭斜率的线性回归,如线性回归1005,表示在较快存储层中执行更好的虚拟机,而具有浅斜率的线性回归,如线性回归1010,表示在较慢存储层中充分执行的虚拟机。

[0203] 图11示出关于图3的存储层340-1、存储层340-2和存储层340-3的图8的专业化信息820。在图11中,专业化信息820示出为矩阵。每一个存储层与每一个资源的交叉点可以指定反映存储层是否提供所述资源(和/或通过多少所述资源)的值。在本发明概念的一些实施例中,如11图中所示,专业化信息830可以使用二进制信息(0与1,是与否,或任何其它对值),其中一个值表明存储层提供所述资源并且另一个值表明存储层不提供所述资源因此举例来说,专业化信息820可以读取为表明存储层1提供高吞吐量和高带宽,存储层2提供高带宽,且存储层3提供高容量。

[0204] 在本发明概念的其它实施例中,专业化信息820中的值不必受限于二进制选项。举

例来说,所述值可以在0到1的范围内,其中0表明存储层不提供所述资源,1表明存储层完整提供所述资源,且两者之间的值可以表示部分支持。因此,比如0.75的值可以表明存储层在一定程度上提供资源,但不与“最好”存储层一样好,而比如0.25的值可以表明存储层提供比“最差”存储层略好的资源(其中“最好”和“最差”为相对术语)。

[0205] 图12示出计算待用于将图3的虚拟机305-1、虚拟机305-2和虚拟机305-3迁移到图3的不同存储层中的得分矩阵。通过专业化信息820与资源矩阵905相乘,可以生成反映每一个层支持每一个虚拟机的程度的得分矩阵1205。举例来说,值被提供给虚拟机1,表明存储层1是55%匹配,存储层2是25%匹配,且存储层3是30%匹配。(注意到所述值并不总计为100%,这是因为每一个值表示每一个存储层将支持虚拟机的程度,而不是特定虚拟机应存储于特定存储层中多少。)可以类似地解释另一虚拟机的项目。

[0206] 一旦产生得分矩阵1205后,图8的层选择825可以为每一个虚拟机选择适当的层。这种选择可以通过计算层和虚拟机的每一个组合的以下值来完成:最终得分=老化因子×历史最终得分+当前性能增益-当前性能代偿。老化因子可以是用于使虚拟机和存储层的组合的最终得分变旧的因子(以防止早期选择过度偏置结果)。当前性能增益可以是根据虚拟机和层的组合的图12的得分矩阵1205绘制的值。当前性能代偿表示从一个存储层迁移数据到另一个存储层的成本。

[0207] 应注意,由于虚拟机可以驻存于除含有图1的存储池135的服务器以外的不同服务器上,当前性能代偿仅反映移动数据本身的成本,且不需要将对虚拟机操作的任何影响作为考虑因素。换句话说,由于虚拟机驻存于图1的虚拟机服务器110-1上,所以托管虚拟机的服务器不需要有处理能力来执行数据迁移:所有处理是在图1的数据中心服务器130上完成的。当前性能代偿可以通过用需要迁移的虚拟机文件大小除以源存储层的更小读取输送率和目标存储层的写入输送率来计算。换句话说,比较从源存储层读取数据所需的时间和将所述数据写入目标存储层所需的时间,且选择更大值作为执行迁移所需的时间,这是因为无论哪个较慢(读取源存储层或写入目标存储层)都表示数据迁移中的瓶颈。

[0208] 给定虚拟机和存储层的每一个组合的最终得分,层选择模块825可以容易地为每一个虚拟机选择存储层以通过确认虚拟机到存储层的分配来全局优化系统105的整体性能,所述分配提供对图1的系统105的性能的整体最大改进,从而将性能改进和性能降低(通过将虚拟机迁移到不太强大的存储层)以及迁移成本作为考虑因素。

[0209] 如上文提示,针对所有时间执行单一全局优化假设每一个虚拟机的工作负载不改变,且虚拟机集群自身不改变(没有添加新的虚拟机且不移除现有的虚拟机)。实际上,这些情况是不真实的。因此,如果图1的系统105仅优化一次,随时间推移,虚拟机到存储层的分配可能不再是最优的。这种问题的解决方案是以规律间隔执行优化,而不是仅一次。但在连续基础上执行优化不是高效的:例如,虚拟机I/O命令将必须在连续基础上经历增加的延迟,意味着无虚拟机会接收其存储层的全部益处。因此,并非执行连续优化,可以在定期基础上执行优化。为此,引入迁移期和监测期的概念。

[0210] 监测期表示图3的I/O滤波器325可以将延迟增加到一些I/O命令中以确定虚拟机现在可以在不同存储层中如何执行的时间点。当监测期开始时,图3的I/O滤波器325可以将延迟增加到所选择数量的I/O命令中。这些数量的I/O命令可以预设为固定数量,或其可以设定为在先前监测期期间发出的I/O命令数量的百分比(比如,1%),或这两个值(比如,这

两个值中的更大值)。监测期可以以规律间隔,例如每15分钟、或每小时或每天出现,但本发明概念的实施例可以支持不规律的监测期。应注意不存在预测在监测期开始时虚拟机的实际工作负载将是什么的简单方式。因此,触发图3的I/O滤波器325以在监测期开始处将延迟增加到I/O命令中有可能将选择随机的I/O命令集且提供图7的有意义的性能数据720。一旦已产生虚拟机在给定监测期中的图7的性能数据720,图3的自动分层后台程序330可以将图7的性能数据720传送到图3的自动分层控制器335。替代地,图3的自动分层后台程序330可以等待直到迁移期开始并在当时传送图7的性能数据720。

[0211] 迁移期表示图3的自动分层控制器335可以确定虚拟机是否应从一个存储层迁移到另一个存储层的时间点。如同监测期,迁移期可设定以规律或不规律时间间隔,例如每天或每周出现。当监测期出现时,图3的自动分层控制器335可以分析图7的性能数据720,产生图9的资源矩阵905,计算最终得分,选择用于每一个虚拟机的存储层,以及必要时迁移虚拟机数据。

[0212] 图13示出监测期与迁移期之间的关系。在图13中,示出迁移期1305和监测期1310。如图13中可见,迁移期1305通常比监测期1310更长,但本发明概念的实施例可以具有与监测期1310一样长或更短的迁移期1305。另外,迁移期1305和监测期1310不需要对准,但其可以对准。

[0213] 图14示出根据本发明概念的实施例的用于图3的I/O滤波器325将延迟增加到图7的输入/输出(I/O)命令705以产生图7的性能数据3615的实例过程的流程图。在图14中,在块1405处,图3的I/O滤波器325可以确定图13的新监测期1310是否已开始。如果未开始,那么图3的I/O滤波器325可以等待。否则,在块1410处,图3的I/O滤波器325可以接收来自图3的应用程序310的图7的一些I/O命令705。图3的I/O滤波器325以这种方式选择多少命令来处理可以根据系统配置改变。在块1415处,图3的I/O滤波器325可以将图7的延迟710增加到图7的I/O命令705中。在块1420处,图3的I/O滤波器325可以将命令传送到图1的数据中心服务器130。在块1425处,图3的I/O滤波器325可以接收来自图1的数据中心服务器130的图7的响应715。在块1418处,图3的I/O滤波器325可以将图7的响应715传递到图3的应用程序310。最后,在块1430处,图3的I/O滤波器325可以使用虚拟机的平均I/O延迟计算图7的性能数据720。

[0214] 图15A到图15B示出根据本发明概念的实施例的用于图3的自动分层控制器335管理虚拟机在存储层中的迁移的实例过程的流程图。在图15A中,在块1505处,图3的自动分层控制器335可以确定图13的新监测期1310是否已开始。如果未开始,那么图3的自动分层控制器335可以等待。否则,在块1510处,图3的自动分层控制器335可以接收来自图3的自动分层后台程序330的图7的性能数据720。在块1515处,图3的自动分层控制器335可以确定图13的新迁移期1305是否已开始。如果未开始,那么图3的自动分层控制器335可以等待。

[0215] 否则,在块1520处,图8的资源矩阵生成器810可以产生图9的资源矩阵905(在其它组件中使用图8的线性回归模块805和图8的度量转换器815)。在块1525处,层选择模块825可以使用资源矩阵905将虚拟机分配到存储层。在块1530处,图3的自动分层控制器335可以检查以查看虚拟机中的任一个是否从一个存储层迁移到另一个存储层。如果是,则在块1535处,迁移模块830可以将图5的虚拟机文件505从一个存储层迁移到另一个存储层。

[0216] 图16示出根据本发明概念的实施例的用于图3的自动分层控制器335以产生图9的

资源矩阵905的实例过程的流程图。在图16中,在块1605处,图8的线性回归模块805可以对图7的性能数据720执行线性回归以产生估计虚拟机将在每一个存储层中执行的程度的方程式。在块1610处,图3的自动分层控制器335可以使用图10的线性回归1005和线性回归1010以估计每一个虚拟机将在每一个存储层中执行的程度。在块1615处,资源矩阵生成器810可以根据每一个资源生成虚拟机在每一个存储层中的性能度量。如块1620中所示,这可以包含使用度量转换器815将一个度量(例如吞吐量)转换成另一个度量(例如带宽)。块1620可以省略,如由虚线1625所示。

[0217] 如从以上论述中可见,本发明概念的实施例提供比常规分层和缓存解决方案显著的优势。第一,本发明概念的实施例支持考虑存储层的多个资源。另外,由于资源数量可以改变,所以资源数量可以设定为1,从而提供与常规HDD存储池一起使用的后向兼容性。第二,本发明概念的实施例对快闪存储系统的较快性能做出解释,忽略常规分层和缓存解决方案。第三,由于现有虚拟机集群的改变和虚拟机工作负载(例如工作负载峰)出现,本发明概念的实施例支持寻找全局优化并随时间推移调节所述全局优化。这与常规解决方案相当,其通常为个人虚拟机寻找本地解决方案而不是整体上优化系统。

[0218] 在图14到图16中,示出本发明概念的一些实施例。但本领域技术人员将认识到,通过改变块的次序、通过省略块,或通过包含未在附图中示出的链路,本发明概念的其它实施例也是可能的。无论是否明确地描述,流程图的所有此类变化都被视为本发明概念的实施例。

[0219] 以下论述预期提供其中可以实施本发明概念的某些方面的一个或多个合适机器的简要一般描述。一个或多个机器可以至少部分地通过从例如键盘、鼠标等常规输入装置的输入,以及通过从另一机器接收的指令、与虚拟现实(virtual reality,VR)环境的交互、生物计量反馈或其它输入信号控制。本文所使用的术语“机器”意欲广泛地涵盖单个机器、虚拟机,或以通信方式耦合的机器、虚拟机或一起操作的装置的系统。示范性机器包含计算装置,例如个人计算机、工作站、服务器、便携式计算机、手持式装置、电话、平板计算机等,以及交通装置,例如私人或公共交通工具,例如汽车、火车、出租车等。

[0220] 一个或多个机器可以包含嵌入式控制器,例如,可编程或不可编程逻辑装置或阵列、专用集成模块(Application Specific Integrated Modules,ASIC)、嵌入式计算机、智能卡等。一个或多个机器可以利用例如通过网络接口、调制解调器或其它通信耦合与一个或多个远程机器的一个或多个连接。机器可以借助于物理和/或逻辑网络,例如企业内部网、因特网、局域网、广域网等互连。本领域技术人员将理解,网络通信可以利用各种有线和/或无线短程或远程载波和协议,包含射频(radio frequency,RF)、卫星、微波、电气电子工程师学会(Institute of Electrical and Electronics Engineer,IEEE)802.11、Bluetooth®、光学、红外线、电缆、激光等。

[0221] 可以参考或结合包含功能、过程、数据结构、应用程序等的相关联数据描述本发明概念的实施例,所述相关联数据在通过机器访问时引起机器执行任务或定义抽象数据类型或低级硬件环境。相关联数据可以存储在例如易失性和/或非易失性存储器,例如RAM、ROM等中,或其它存储装置以及其相关联存储介质中,包含硬盘驱动器、软盘、光学存储装置、磁带、快闪存储器、存储棒、数字视频光盘、生物存储装置等。相关联数据可以采用包、串行数据、并行数据、传播信号等形式在包含物理和/或逻辑网络的传输环境上传递,并且可以用

压缩或加密格式使用。相关联数据可以用于分布式环境中,并且本地地和/或远程地存储用于机器访问。

[0222] 本发明概念的实施例可以包含有形的非暂时性机器可读介质,包括可由一个或多个处理器执行的指令,所述指令包括用于执行如本文所描述的发明性概念的元件的指令。

[0223] 上文所描述的方法的各种操作可以通过能够执行所述操作的任何合适构件(例如各种硬件和/或软件组件、模组及/或一个或多个模块)来执行。软件可以包含用于实施逻辑功能的可执行指令的有序列表,且可以任何供指令执行系统、设备或装置使用或与指令执行系统、设备或装置结合的“处理器可读介质”实施,所述指令执行系统、设备或装置如单核处理器或多核处理器或含处理器的系统。

[0224] 结合本文中所公开的实施例而描述的方法或算法和函数的步骤可直接体现在硬件、由处理器执行的软件模块或所述两者的组合中。如果实施于软件中,则可以将功能作为一个或多个指令或代码而存储在有形的非暂时计算机可读介质上或经由计算机可读介质进行传输。软件模块可驻留在随机存取存储器(RAM)、快闪存储器、只读存储器(ROM)、可电编程ROM(EPROM)、可电擦除可编程ROM(EEPROM)、寄存器、硬盘、可装卸磁盘、CD ROM或所属领域中已知的任何其它形式的存储介质中。

[0225] 在参考所说明实施例描述和说明本发明概念的原理之后,将认识到,所说明实施例的布置和细节可以在不脱离此原理的情况下进行修改,并且可以按任何所需方式组合。而且,尽管前述论述集中于具体实施例,但预期其它配置。具体地说,即使本文使用例如“根据本发明概念的实施例”等的表达,但这些词语一般表示参考实施例可能性,且并不意图将本发明概念限制到具体实施例配置。如本文所使用,这些术语可以参考组合到其它实施例中的相同或不同实施例。

[0226] 前述说明性实施例不应被解释为限制其本发明概念。虽然已经描述几个实施例,但本领域技术人员将容易了解,在实质上不脱离本公开的新颖教导和优点的情况下,对这些实施例的许多修改是可能的。因此,所有此类修改意欲包含在如所附权利要求中定义的本发明概念的范围內。

[0227] 本发明概念的实施例可以扩展到以下陈述,而不具有限制性:

[0228] 陈述1. 本发明概念的实施例包含一种系统,所述系统包括:

[0229] 提供多个资源的多个存储装置,所述多个存储装置组织成多个存储层并且存储第一虚拟机的第一数据和第二虚拟机的第二数据;

[0230] 接收器,接收来自所述第一虚拟机的第一输入/输出(I/O)命令、来自所述第二虚拟机的第二I/O命令、模拟所述第一虚拟机在所述多个存储层中的性能的第一性能数据,以及模拟所述第二虚拟机在所述多个存储层中的性能的第二性能数据;

[0231] 传输器,将对于所述第一I/O命令的第一响应传输到所述第一虚拟机并且将对于所述第二I/O命令的第二响应传输到所述第二虚拟机;以及

[0232] 自动分层控制器,选择第一存储层以存储所述第一虚拟机的所述第一数据、选择第二存储层以存储所述第二虚拟机的所述第二数据,并且响应于所述第一性能数据和所述第二性能数据来迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个,

[0233] 其中所述自动分层控制器可操作以选择所述第一存储层以存储所述第一虚拟机

的所述第一数据并且选择所述第二存储层以存储所述第二虚拟机的所述第二数据来优化所有虚拟机在所述多个存储层中的性能,

[0234] 其中所述自动分层控制器可操作以将由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所引起的性能改变和迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个的迁移成本作为考虑因素。

[0235] 陈述2. 本发明概念的实施例包含根据陈述1所述的系统, 其中:

[0236] 所述系统进一步包括处理器; 以及

[0237] 所述自动分层控制器包含在所述处理器上运行的软件。

[0238] 陈述3. 本发明概念的实施例包含根据陈述1所述的系统, 其中所述自动分层控制器包含:

[0239] 层选择模块, 响应于所述第一性能数据和所述第二性能数据选择所述第一虚拟机的所述第一存储层和所述第二虚拟机的所述第二存储层; 以及

[0240] 迁移模块, 迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个。

[0241] 陈述4. 本发明概念的实施例包含根据陈述1所述的系统, 其中所述自动分层控制器包含使用所述第一性能数据和所述第二性能数据生成资源矩阵的资源矩阵生成器, 所述资源矩阵使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联。

[0242] 陈述5. 本发明概念的实施例包含根据陈述4所述的系统, 其中所述自动分层控制器可操作以响应于所述资源矩阵来选择第一存储层以存储所述第一虚拟机的所述第一数据并且选择第二存储层以存储所述第二虚拟机的所述第二数据。

[0243] 陈述6. 本发明概念的实施例包含根据陈述5所述的系统, 其中所述自动分层控制器可操作以响应于所述多个存储层中的每一个的所述资源矩阵和专业化信息来选择第一存储层以存储所述第一虚拟机的所述第一数据并且选择第二存储层以存储所述第二虚拟机的所述第二数据。

[0244] 陈述7. 本发明概念的实施例包含根据陈述6所述的系统, 其中所述多个存储层中的每一个的所述专业化信息确认所述多个存储层中的每一个是否支持所述多个资源中的每一个。

[0245] 陈述8. 本发明概念的实施例包含根据陈述7所述的系统, 其中所述多个资源包含吞吐量、带宽、容量和写入放大。

[0246] 陈述9. 本发明概念的实施例包含根据陈述4所述的系统, 其中所述资源矩阵生成器包含根据所述多个资源中的第二资源将所述第一虚拟机在所述多个存储层中的每一个中的第一性能度量转换成所述第一虚拟机在所述多个存储层中的每一个中的第二性能度量的度量转换器。

[0247] 陈述10. 本发明概念的实施例包含根据陈述9所述的系统, 其中所述第一性能度量包含所述第一虚拟机的第一平均I/O延迟数据。

[0248] 陈述11. 本发明概念的实施例包含根据陈述1所述的系统, 其中:

[0249] 所述自动分层控制器包含对所述第一性能数据执行第一线性回归并且对所述第

二性能数据执行第二线性回归的线性回归模块;以及

[0250] 所述自动分层控制器可操作以响应于所述第一线性回归和所述第二线性回归选择所述第一存储层以存储所述第一虚拟机的所述第一数据且选择所述第二存储层以存储所述第二虚拟机的所述第二数据。

[0251] 陈述12.本发明概念的实施例包含根据陈述11所述的系统,其中所述自动分层控制器可操作以响应于所述第一线性回归和所述第二线性回归,使用所述第一虚拟机和所述第二虚拟机的估计平均I/O延迟来选择所述第一存储层以存储所述第一虚拟机的所述第一数据且选择所述第二存储层以存储所述第二虚拟机的所述第二数据。

[0252] 陈述13.本发明概念的实施例包含根据陈述1所述的系统,其中:

[0253] 所述第一性能数据包含使用第一资源的第一度量性能数据;以及

[0254] 所述自动分层控制器包含将使用所述第一资源的所述第一度量性能数据转换成使用第二资源的第二度量性能数据的度量转换器。

[0255] 陈述14.本发明概念的实施例包含根据陈述1所述的系统,其中所述自动分层控制器可操作以选择第一存储层来存储所述第一虚拟机的所述第一数据、选择第二存储层来存储所述第二虚拟机的所述第二数据、并且每个迁移期一次响应于所述第一性能数据和所述第二性能数据迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个。

[0256] 陈述15.本发明概念的实施例包含根据陈述14所述的系统,其中:

[0257] 所述第一性能数据包含来自所述迁移期内的多个监测期的第一性能数据;以及

[0258] 所述第二性能数据包含来自所述迁移期内的所述多个监测期的第二性能数据。

[0259] 陈述16.本发明概念的实施例包含根据陈述1所述的系统,所述系统进一步包括:

[0260] 第一虚拟机服务器,存储所述第一虚拟机;以及

[0261] 第二虚拟机服务器,存储所述第二虚拟机。

[0262] 陈述17.本发明概念的实施例包含根据陈述16所述的系统,其中所述第二虚拟机服务器是所述第一虚拟机服务器。

[0263] 陈述18.本发明概念的实施例包含根据陈述16所述的系统,其中所述第一虚拟机服务器包含产生模拟所述第一虚拟机在所述多个存储层中的性能的所述第一性能数据的I/O滤波器。

[0264] 陈述19.本发明概念的实施例包含根据陈述18所述的系统,其中所述I/O滤波器可操作以将延迟增加到来自所述第一虚拟机的多个I/O命令、包含所述第一I/O命令的多个I/O命令中,并且响应于所述增加的延迟确定所述第一虚拟机的平均I/O延迟。

[0265] 陈述20.本发明概念的实施例包含根据陈述19所述的系统,其中:

[0266] 所述I/O滤波器进一步可操作以将多个延迟增加到来自所述第一虚拟机的多个I/O命令中,并且响应于所述增加的延迟确定所述第一虚拟机的多个平均I/O延迟;以及

[0267] 所述自动分层控制器包含对所述第一虚拟机的所述多个平均I/O延迟执行第一线性回归的线性回归模块。

[0268] 陈述21.本发明概念的实施例包含根据陈述18所述的系统,其中所述第一虚拟机服务器进一步包含从所述I/O滤波器收集多个虚拟机的性能数据并且将所汇集的性能数据传输到所述自动分层控制器的自动分层后台程序。

[0269] 陈述22.本发明概念的实施例包含根据陈述16所述的系统,其中:

[0270] 所述第一虚拟机服务器不包含存储所述第一虚拟机的第一数据中的任一种的第一本地缓存;以及

[0271] 所述第二虚拟机服务器不包含存储所述第二虚拟机的第二数据中的任一种的第二本地缓存。

[0272] 陈述23.本发明概念的实施例包含根据陈述1所述的系统,其中由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所造成的性能改变包含所述第一虚拟机的性能改进。

[0273] 陈述24.本发明概念的实施例包含根据陈述1所述的系统,其中由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所造成的性能改变包含所述第一虚拟机的性能下降。

[0274] 陈述25.本发明概念的实施例包含根据陈述1所述的系统,其中迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个的迁移成本包含迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所需要的迁移时间。

[0275] 陈述26.本发明概念的实施例包含根据陈述1所述的系统,其中所述第二存储层是所述第一存储层。

[0276] 陈述27.本发明概念的实施例包含根据陈述1所述的系统,其中所述多个存储层中的一个中的多个存储装置是均质的。

[0277] 陈述28.本发明概念的实施例包含一种方法,所述方法包括:

[0278] 接收模拟第一虚拟机在多个存储层中的性能的第一性能数据;

[0279] 接收模拟第二虚拟机在所述多个存储层中的性能的第二性能数据;

[0280] 使用所述第一性能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵;

[0281] 使用所述资源矩阵选择所述多个存储层中的第一存储层以存储所述第一虚拟机的第一数据;

[0282] 使用所述资源矩阵选择所述多个存储层中的第二存储层以存储所述第二虚拟机的第二数据;以及

[0283] 迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个,

[0284] 其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据使所有虚拟机在所述多个存储层中的性能优化,

[0285] 其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所引起的性能改变作为考虑因素,以及

[0286] 其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个的迁移成本作为考虑因素。

[0287] 陈述29. 本发明概念的实施例包含根据陈述28所述的方法, 其中所述第二存储层是所述第一存储层。

[0288] 陈述30. 本发明概念的实施例包含根据陈述28所述的方法, 其中接收第一性能数据、接收第二性能数据、生成资源矩阵、选择第一存储层和选择第二存储层是每迁移期执行一次。

[0289] 陈述31. 本发明概念的实施例包含根据陈述30所述的方法, 其中接收第一性能数据和接收第二性能数据是每监测期执行一次。

[0290] 陈述32. 本发明概念的实施例包含根据陈述31所述的方法, 其中所述迁移期包含多个监测期。

[0291] 陈述33. 本发明概念的实施例包含根据陈述28所述的方法, 其中:

[0292] 使用所述资源矩阵选择所述多个存储层中的第一存储层以存储所述第一虚拟机的第一数据包含使用所述资源矩阵和所述多个存储层中的每一个的所述专业化信息选择所述多个存储层中的所述第一存储层以存储所述第一虚拟机的第一数据; 以及

[0293] 使用所述资源矩阵选择所述多个存储层中的第二存储层以存储所述第二虚拟机的第二数据包含使用所述资源矩阵和所述多个存储层中的每一个的所述专业化信息选择所述多个存储层中的所述第二存储层以存储所述第二虚拟机的第二数据。

[0294] 陈述34. 本发明概念的实施例包含根据陈述33所述的方法, 其中所述多个存储层中的每一个的所述专业化信息确认所述多个存储层中的每一个是否支持所述多个资源中的每一个。

[0295] 陈述35. 本发明概念的实施例包含根据陈述34所述的方法, 其中所述多个资源包含吞吐量、带宽、容量和写入放大。

[0296] 陈述36. 本发明概念的实施例包含根据陈述34所述的方法, 其中模拟所述第一虚拟机在所述多个存储层中的性能的所述第一性能数据包含使所述第一虚拟机的平均输入/输出 (I/O) 延迟与增加的延迟相关联的多个数据点。

[0297] 陈述37. 本发明概念的实施例包含根据陈述36所述的方法, 其中使用所述第一性能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵包含对使所述第一虚拟机的平均I/O延迟与增加的延迟相关联的所述多个数据点执行线性回归。

[0298] 陈述38. 本发明概念的实施例包含根据陈述37所述的方法, 其中使用所述第一性能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵进一步包含使用所述线性回归估计所述第一虚拟机在所述多个存储层中的每一个中的性能。

[0299] 陈述39. 本发明概念的实施例包含根据陈述37所述的方法, 其中使用所述第一性能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵进一步包含根据

所述多个资源中的每一个确定所述第一虚拟机在所述多个存储层中的每一个中的性能度量。

[0300] 陈述40. 本发明概念的实施例包含根据陈述39所述的方法, 其中根据所述多个资源中的每一个确定所述第一虚拟机在所述多个存储层中的每一个中的性能度量包含将根据所述多个资源中的第一资源的所述第一虚拟机在所述多个存储层中的每一个中的第一性能度量转换成根据所述多个资源中的第二资源的所述第一虚拟机在所述多个存储层中的每一个中的第二性能度量。

[0301] 陈述41. 本发明概念的实施例包含根据陈述28所述的方法, 其中接收模拟第一虚拟机在多个存储层中的性能的第一性能数据包含:

[0302] 在I/O滤波器处接收来自所述虚拟机的多个第一I/O命令;

[0303] 将第一延迟增加到所述多个第一I/O命令中; 以及

[0304] 针对所述多个第一I/O命令计算所述虚拟机的第一平均I/O延迟。

[0305] 陈述42. 本发明概念的实施例包含根据陈述41所述的方法, 其中:

[0306] 将第一延迟增加到所述多个第一I/O命令中包含将多个延迟增加到所述多个第一I/O命令中;

[0307] 针对所述多个第一I/O命令计算所述虚拟机的第一平均I/O延迟包含针对所述多个第一I/O命令计算所述虚拟机的多个平均I/O延迟; 以及

[0308] 所述方法进一步包括对所述多个平均I/O延迟执行线性回归。

[0309] 陈述43. 本发明概念的实施例包含根据陈述28所述的方法, 其中由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所造成的性能改变包含所述第一虚拟机的性能改进。

[0310] 陈述44. 本发明概念的实施例包含根据陈述28所述的方法, 其中由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所造成的性能改变包含所述第一虚拟机的性能下降。

[0311] 陈述45. 本发明概念的实施例包含根据陈述28所述的方法, 其中迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个的迁移成本包含迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所需要的迁移时间。

[0312] 陈述46. 本发明概念的实施例包含一种装置, 包括非暂时性存储介质, 所述非暂时性存储介质具有存储于其上的在由机器执行时引起以下操作的指令:

[0313] 接收模拟第一虚拟机在多个存储层中的性能的第一性能数据;

[0314] 接收模拟第二虚拟机在所述多个存储层中的性能的第二性能数据;

[0315] 使用所述第一性能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵;

[0316] 使用所述资源矩阵选择所述多个存储层中的第一存储层以存储所述第一虚拟机的第一数据;

[0317] 使用所述资源矩阵选择所述多个存储层中的第二存储层以存储所述第二虚拟机

的第二数据;以及

[0318] 迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个,

[0319] 其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据使所有虚拟机在所述多个存储层中的性能优化,

[0320] 其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所引起的性能改变作为考虑因素,以及

[0321] 其中选择所述第一存储层以存储所述第一虚拟机的所述第一数据和选择所述第二存储层以存储所述第二虚拟机的所述第二数据将迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个的迁移成本作为考虑因素。

[0322] 陈述47.本发明概念的实施例包含根据陈述46所述的装置,其中所述第二存储层是所述第一存储层。

[0323] 陈述48.本发明概念的实施例包含根据陈述46所述的装置,其中接收第一性能数据、接收第二性能数据、生成资源矩阵、选择第一存储层和选择第二存储层是每迁移期执行一次。

[0324] 陈述49.本发明概念的实施例包含根据陈述48所述的装置,其中接收第一性能数据和接收第二性能数据是每监测期执行一次。

[0325] 陈述50.本发明概念的实施例包含根据陈述49所述的装置,其中所述迁移期包含多个监测期。

[0326] 陈述51.本发明概念的实施例包含根据陈述46所述的装置,其中:

[0327] 使用所述资源矩阵选择所述多个存储层中的第一存储层以存储所述第一虚拟机的第一数据包含使用所述资源矩阵和所述多个存储层中的每一个的专业化信息选择所述多个存储层中的所述第一存储层以存储所述第一虚拟机的第一数据;以及

[0328] 使用所述资源矩阵选择所述多个存储层中的第二存储层以存储所述第二虚拟机的第二数据包含使用所述资源矩阵和所述多个存储层中的每一个的所述专业化信息选择所述多个存储层中的所述第二存储层以存储所述第二虚拟机的第二数据。

[0329] 陈述52.本发明概念的实施例包含根据陈述51所述的装置,其中所述多个存储层中的每一个的所述专业化信息确认所述多个存储层中的每一个是否支持所述多个资源中的每一个。

[0330] 陈述53.本发明概念的实施例包含根据陈述52所述的装置,其中所述多个资源包含吞吐量、带宽、容量和写入放大。

[0331] 陈述54.本发明概念的实施例包含根据陈述52所述的装置,其中模拟所述第一虚拟机在所述多个存储层中的性能的所述第一性能数据包含使所述第一虚拟机的平均输入/输出(I/O)延迟与增加的延迟相关联的多个数据点。

[0332] 陈述55.本发明概念的实施例包含根据陈述54所述的装置,其中使用所述第一性

能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵包含对使所述第一虚拟机的平均I/O延迟与增加的延迟相关联的所述多个数据点执行线性回归。

[0333] 陈述56. 本发明概念的实施例包含根据陈述55所述的装置, 其中使用所述第一性能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵进一步包含使用所述线性回归估计所述第一虚拟机在所述多个存储层中的每一个中的性能。

[0334] 陈述57. 本发明概念的实施例包含根据陈述55所述的装置, 其中使用所述第一性能数据和所述第二性能数据生成使所述第一虚拟机和所述第二虚拟机与所述多个存储层中的每一个和所述多个存储层中的每一个的多个资源相关联的资源矩阵进一步包含根据所述多个资源中的每一个确定所述第一虚拟机在所述多个存储层中的每一个中的性能度量。

[0335] 陈述58. 本发明概念的实施例包含根据陈述57所述的装置, 其中根据所述多个资源中的每一个确定所述第一虚拟机在所述多个存储层中的每一个中的性能度量包含将根据所述多个资源中的第一资源的所述第一虚拟机在所述多个存储层中的每一个中的第一性能度量转换成根据所述多个资源中的第二资源的所述第一虚拟机在所述多个存储层中的每一个中的第二性能度量。

[0336] 陈述59. 本发明概念的实施例包含根据陈述46所述的装置, 其中接收模拟第一虚拟机在多个存储层中的性能的第一性能数据包含:

[0337] 在I/O滤波器处接收来自所述虚拟机的第一I/O命令;

[0338] 将第一延迟增加到所述第一I/O命令中; 以及

[0339] 针对所述第一I/O命令计算所述虚拟机的第一平均I/O延迟。

[0340] 陈述60. 本发明概念的实施例包含根据陈述59所述的装置, 其中:

[0341] 将第一延迟增加到所述多个第一I/O命令中包含将多个延迟增加到所述多个第一I/O命令中;

[0342] 针对所述多个第一I/O命令计算所述虚拟机的第一平均I/O延迟包含针对所述多个第一I/O命令计算所述虚拟机的多个平均I/O延迟; 以及

[0343] 所述非暂时性存储介质具有存储于其上的当由机器执行时引起执行对所述多个平均I/O延迟的线性回归的其它指令。

[0344] 陈述61. 本发明概念的实施例包含根据陈述46所述的装置, 其中由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所造成的性能改变包含所述第一虚拟机的性能改进。

[0345] 陈述62. 本发明概念的实施例包含根据陈述46所述的装置, 其中由迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所造成的性能改变包含所述第一虚拟机的性能下降。

[0346] 陈述63. 本发明概念的实施例包含根据陈述46所述的装置, 其中迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的至少一个的迁移成本包含迁移所述第一虚拟机的所述第一数据到所述第一存储层或所述第二虚拟机的所述第二数据到所述第二存储层中的所述至少一个所需要的迁移

时间。

[0347] 因此,考虑到对本文所描述的实施例的各种置换,此详细描述和随附材料预期仅为说明性的,并且不应被视为限制本发明概念的范围。因此,本发明概念所要求保护的是落入所附权利要求书以及其等效物的范围和精神内的所有此类修改。

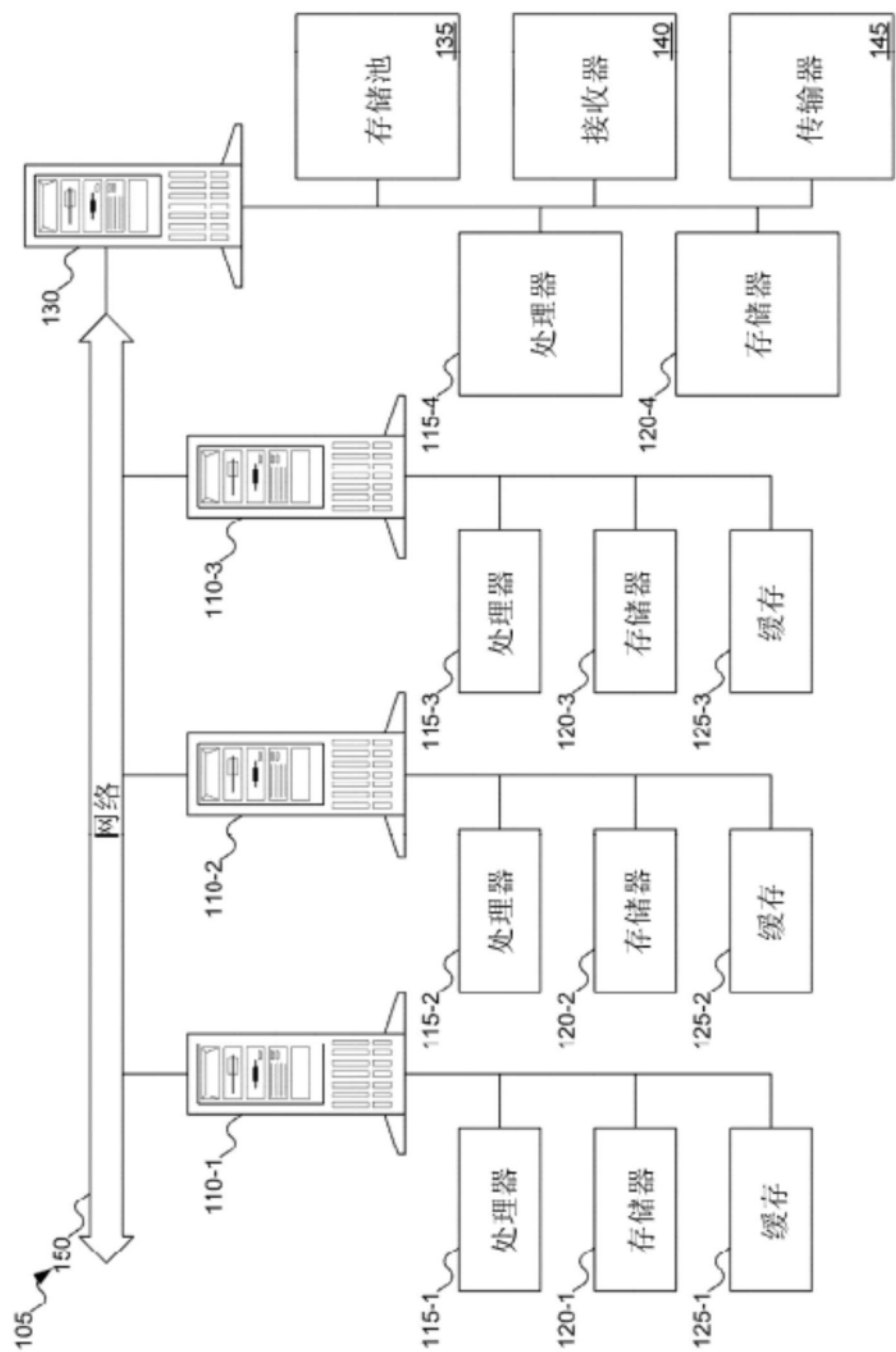


图1

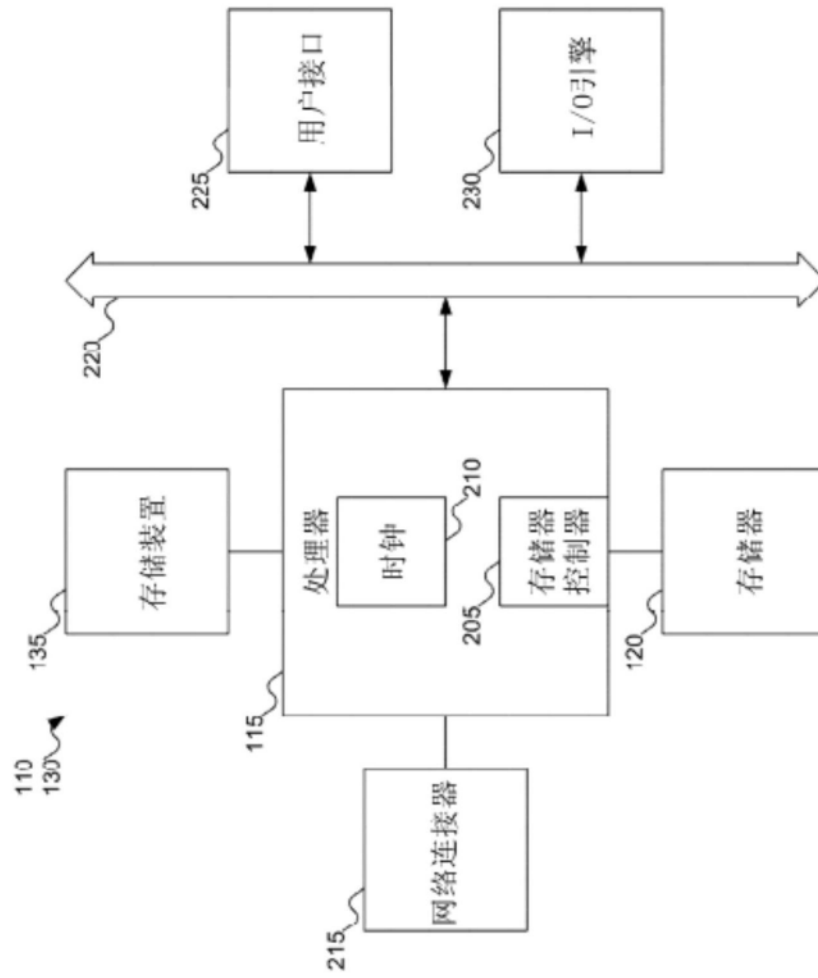


图2

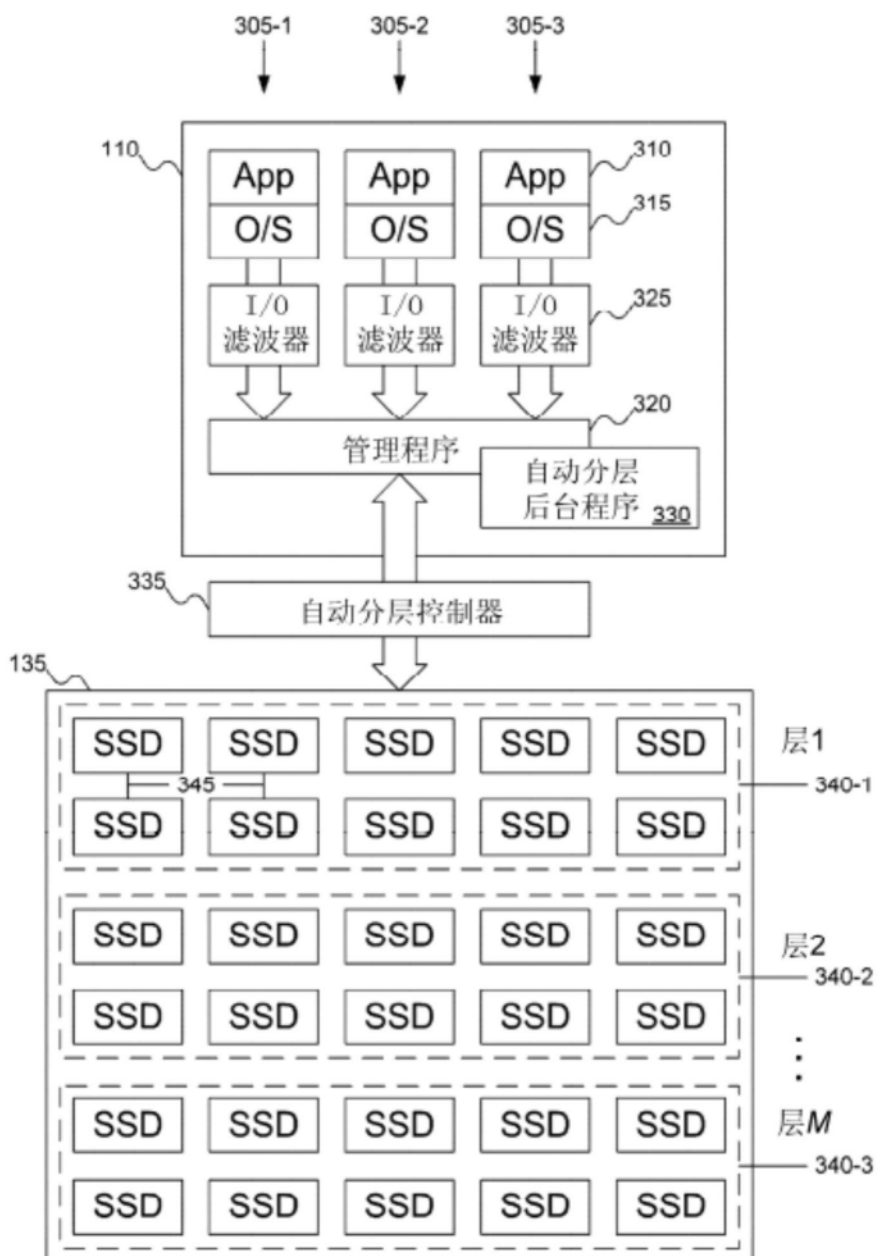


图3

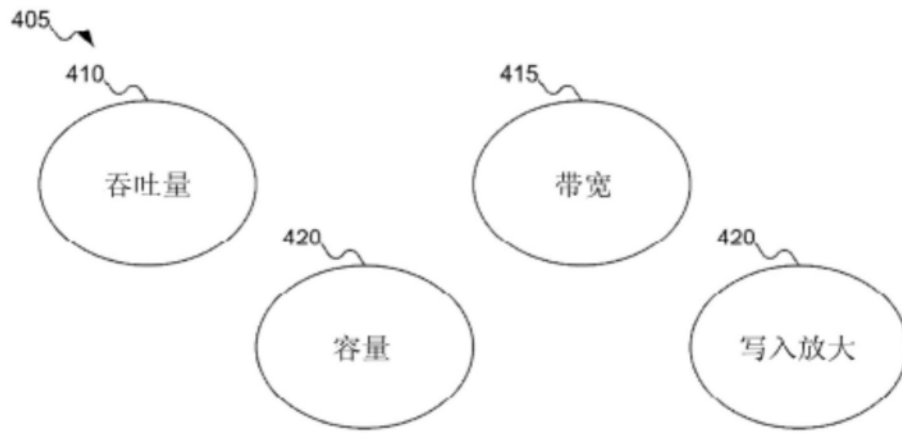


图4

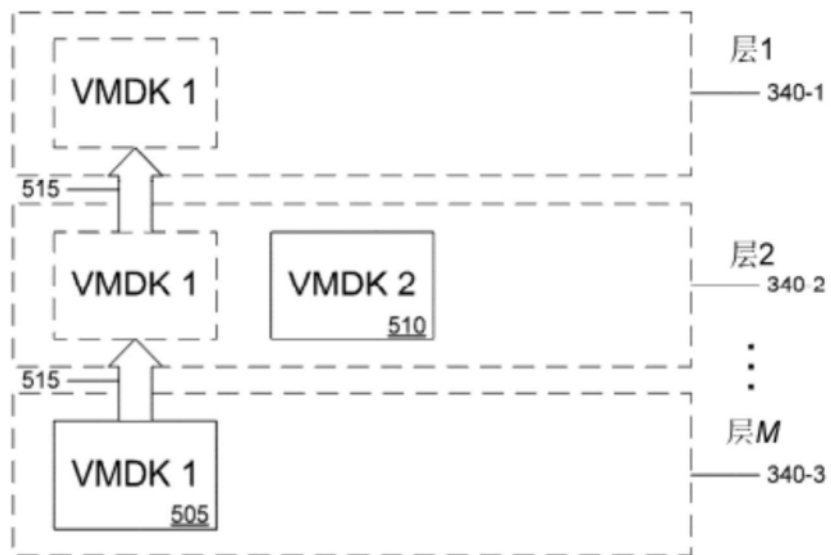


图5

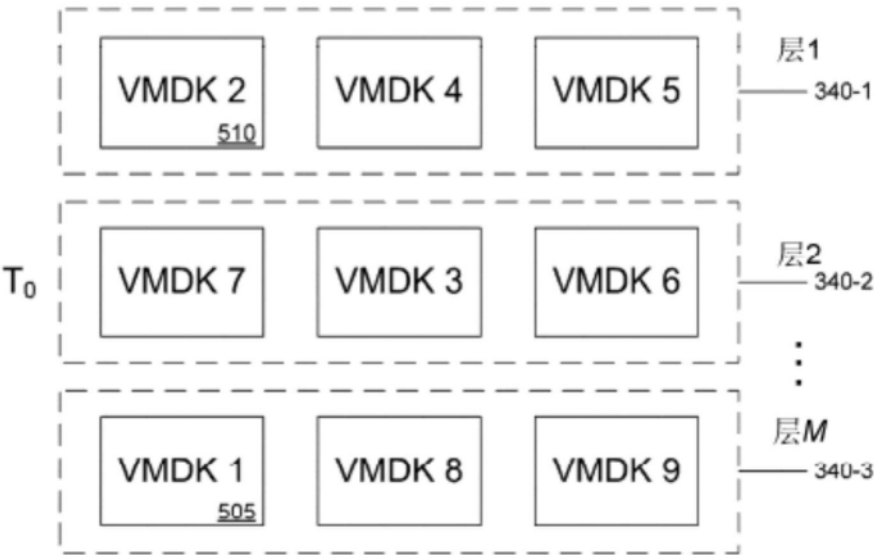


图6A

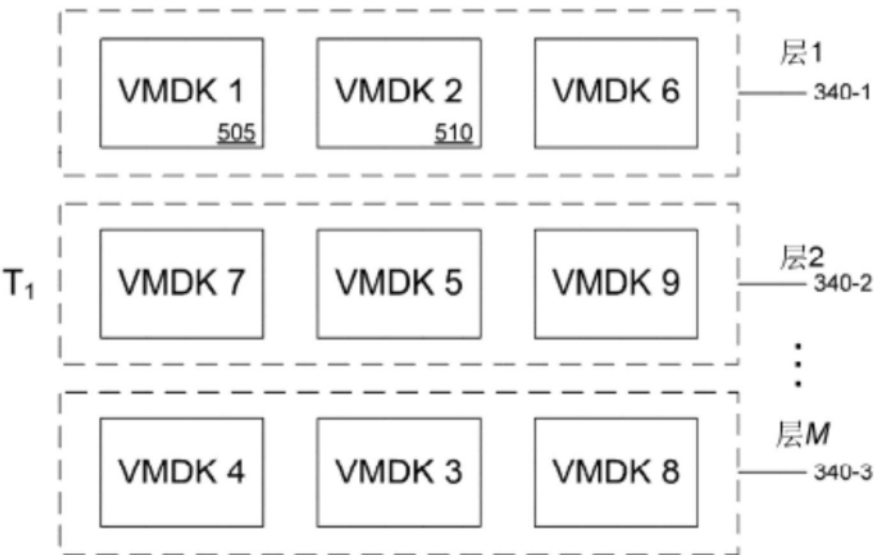


图6B

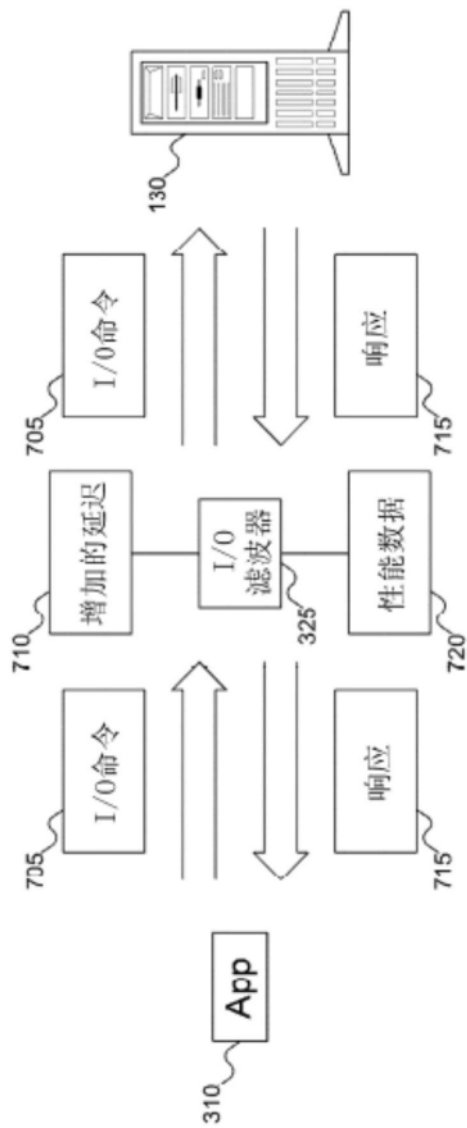


图7

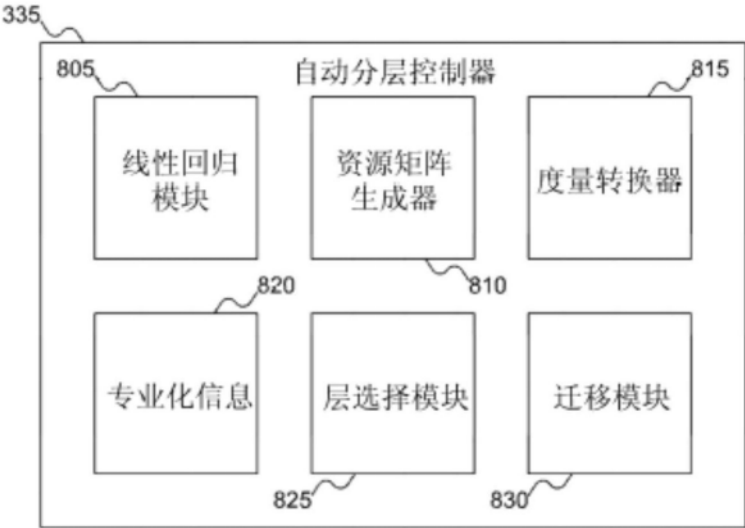


图8

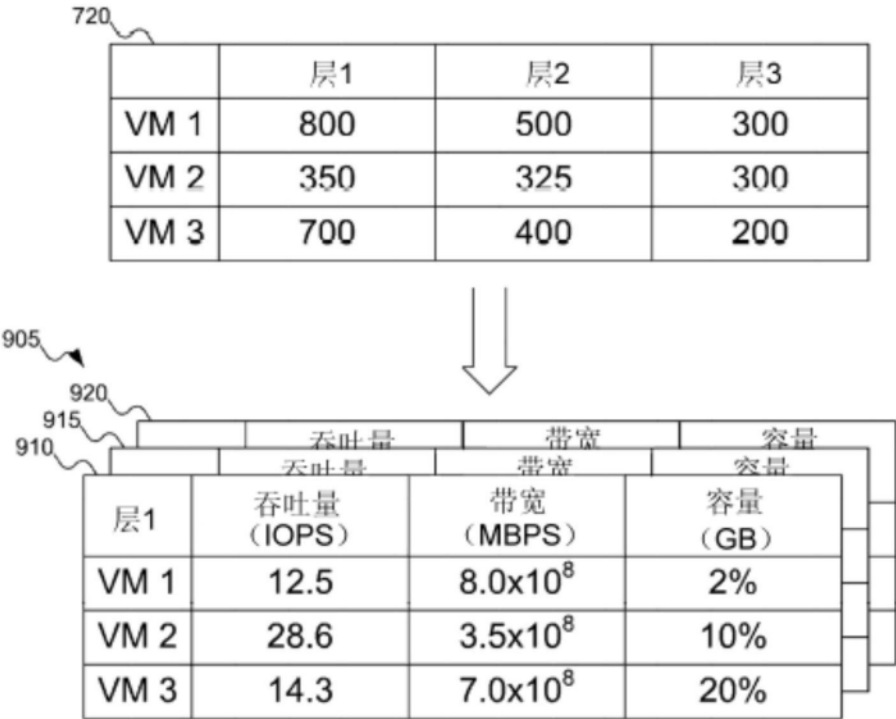


图9

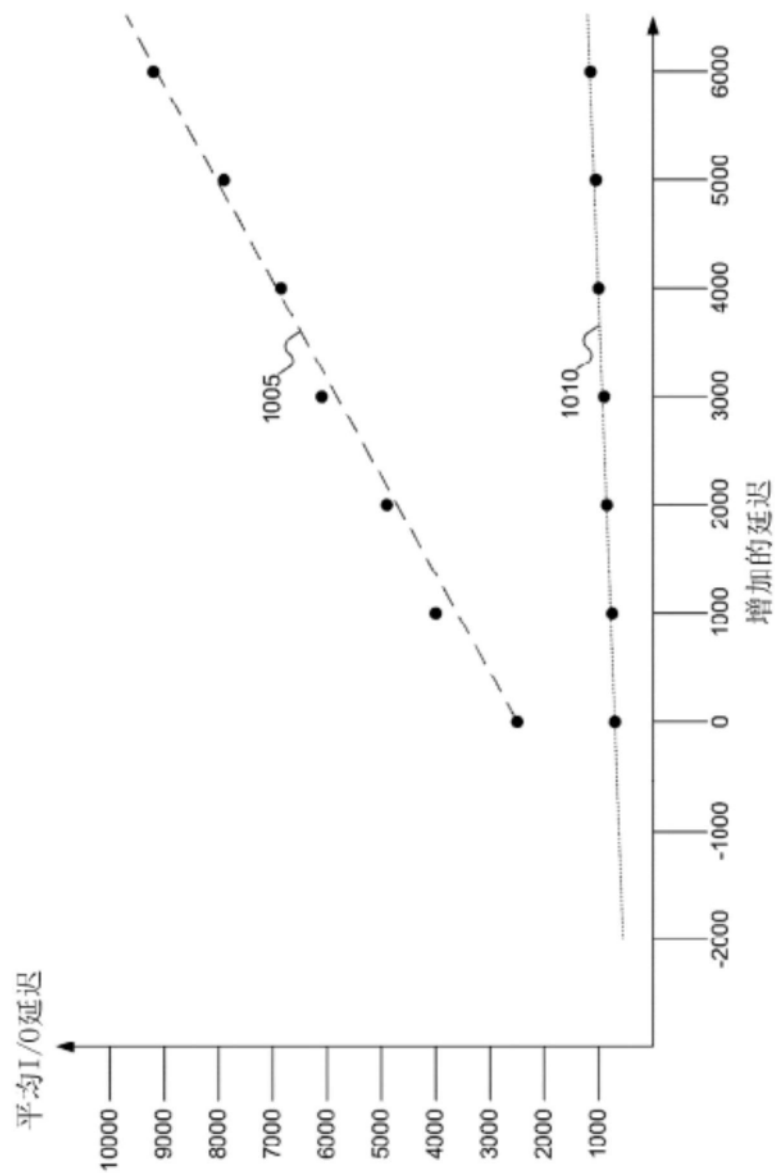


图10

820

专业化信息			
	吞吐量 (IOPS)	带宽 (MBPS)	存储大小 (GB)
层1	1	1	0
层2	0	1	0
层3	0	0	1

图11

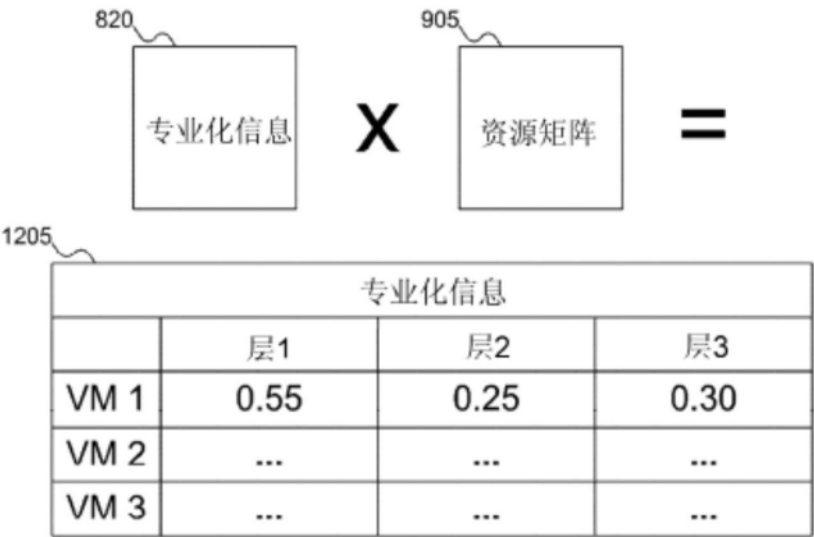


图12

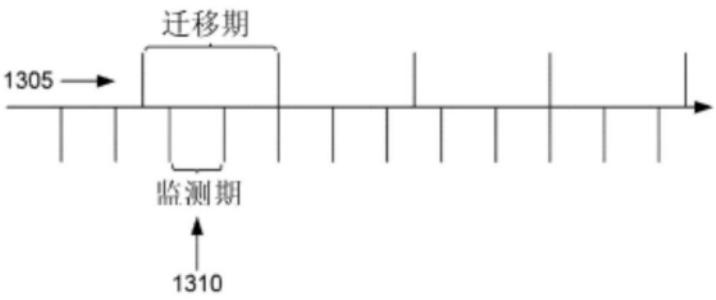


图13

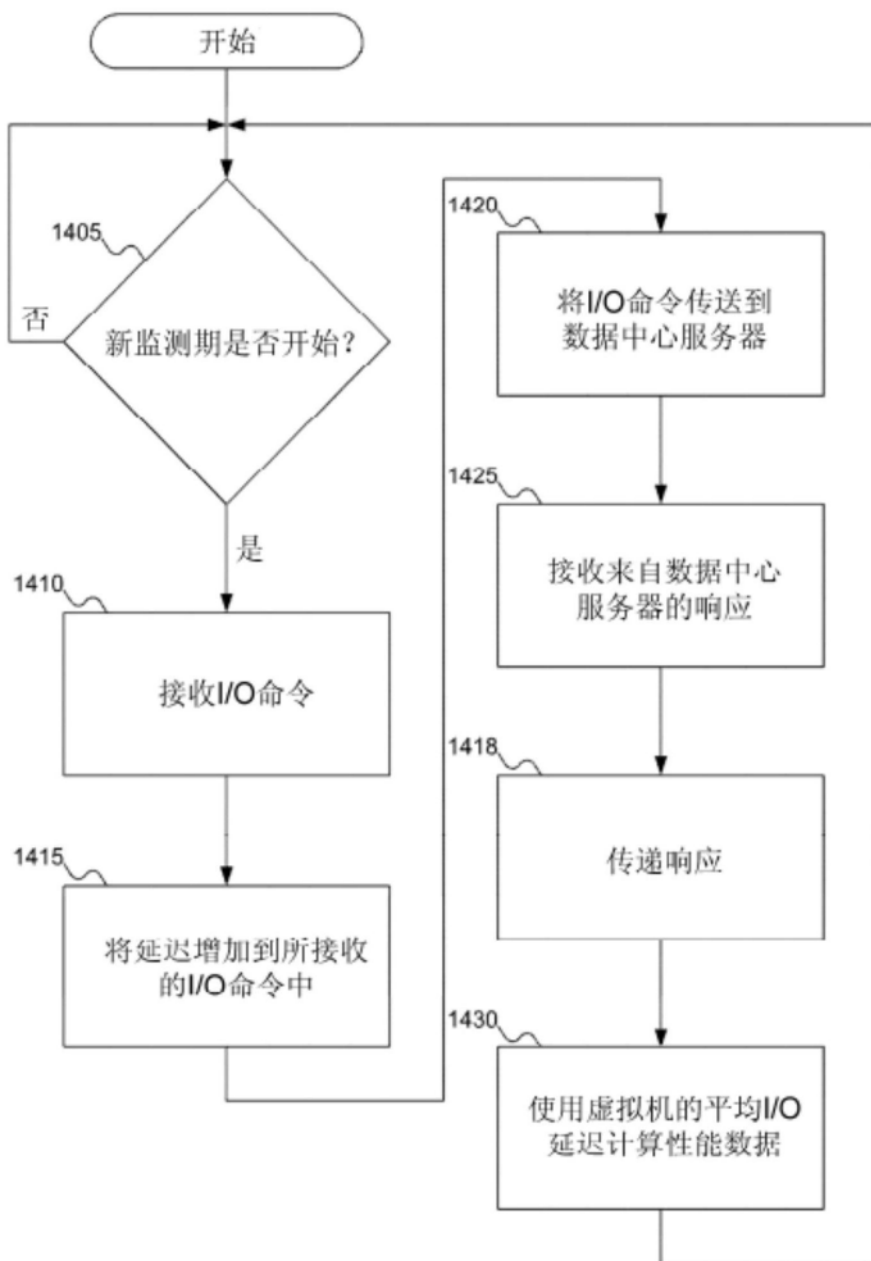


图14

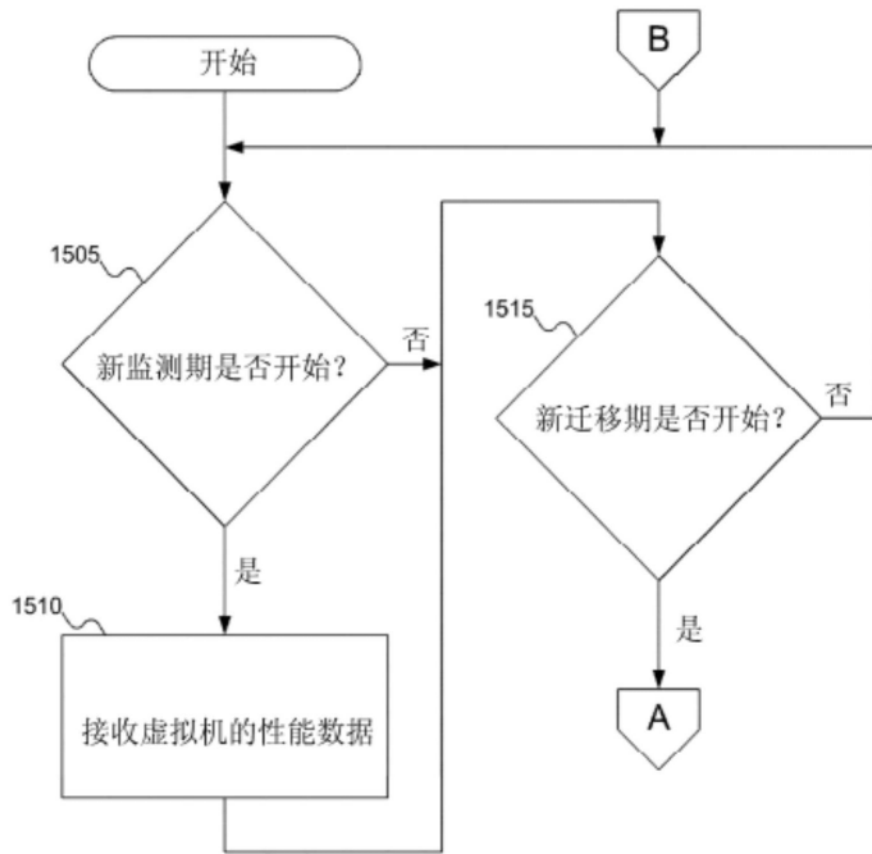


图15A

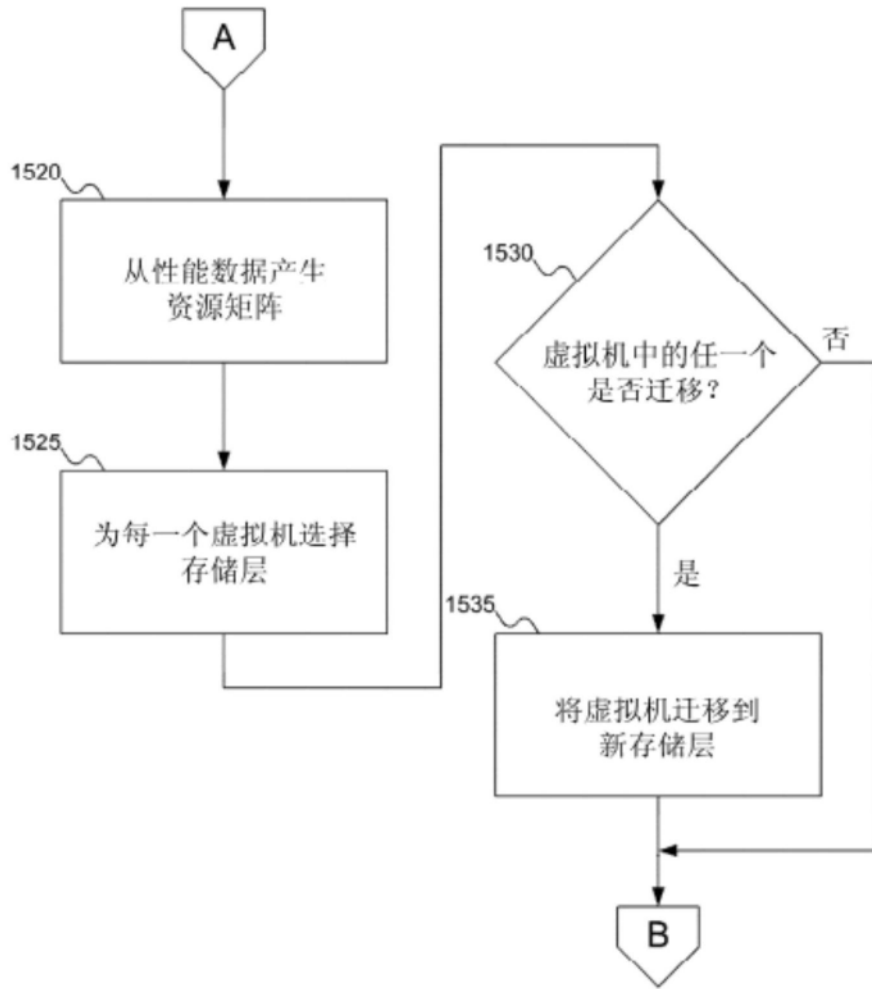


图15B

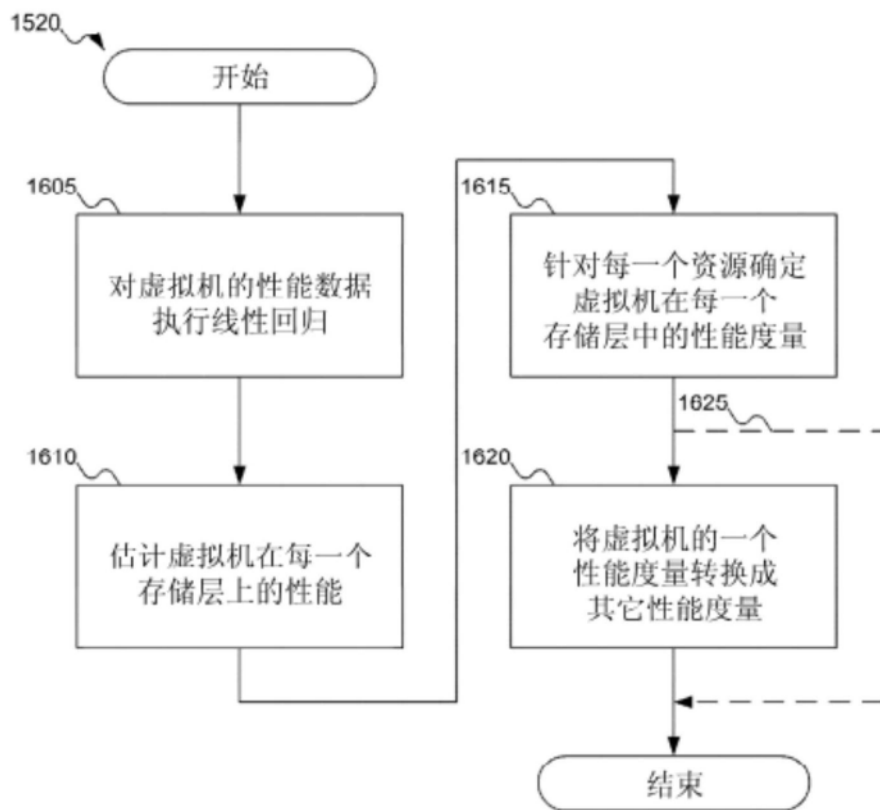


图16