Placement of Virtual Containers on NUMA systems:

A Practical and Comprehensive Model

论文背景综述

这篇论文是在数据中心，解决容器的虚拟核到物理核的映射问题，来获得最高性能，从而降低数据中心能耗。本文讨论两个背景方向，一是NUMA系统的工作配置，二是数据中心的节能技术。

一、NUMA系统的工作配置

NUMA系统上的工作配置的研究已有数年。多核系统中，共享资源可能会导致在不同和上运行的应用程序发生争用，从而导致性能和能效降低。不少人在不同的共享资源不同的角度上提出了尽可能地避免或减少争用带来的性能降低的方法。 Allan S等证明了硬件多线程处理器的性能受作业调度的影响，并提出了一种机制，提高了SMT体系结构的性能。Federova等提出了一种操作系统调度算法，应用于应用程序争用共享高速缓存时，保证它的公平分配和程序运行，也改善了性能隔离。，Merker等提出了任务活动向量，根据任务活动向量进行应用迁移和协同调度，提高应用程序的性能。由数据密集应用程序的内存流量引起的内存控制器和互连的拥塞会对性能造成很大损害，Mohammad等提出了一种内存配置算法解决该问题。这些都为之后探索NUMA系统上预测工作负载的性能做了铺垫。

虚拟核到物理核的映射方法不同会影响共享资源的使用，从而影响系统的性能。相反，可以通过预测NUMA系统的性能，可以找到合适的虚拟核到物理核的映射配置。

Dwyer等提出了一种自动化的模型构建方法，将系统上可用的性能硬件事件输入到各种机器学习模型中。该模型预测了一个相当简单的结果，当目标工作负载与干扰工作负载共同调度时性能下降。该模型没有研究不同配置下的性能，同时其预测的性能准确率也不高。

Daniel等构建了内存并行负载的性能预测系统Pandia，Pandia不仅可以准确地预测多核NUMA机器上不同工作负载配置的性能，还可以预测应用程序如何使用不同数量的线程执行。但是，为了做出预测，Pandia需要先进行具有不同线程数的六次运行，并做性能观察，这一点在实际中难以进行，因为大多数实际应用程序无法轻松地按需重新配置其线程数。尽管解决了以前工作的局限性，但从根本上说，Pandia仍然依赖于特定机器的建模方法，该方法无法轻松地将结果应用到其他系统。 Pandia在一组特定机器的方程中捕获了导致性能的因素，例如缓存争用，通信延迟和负载平衡。如果模型必须适应另一台机器，则必须手动重新制定方程式。

Zhang等提出了SMiTe方法，该方法可以在多核SMT处理器上实现精确的性能干扰预测。由于应用程序在不同共享资源之间的争用特性几乎没有相关性，因此设计了一组标尺来以解耦的方式量化应用程序的敏感性和争用性。 之后建立一个回归模型，该模型结合了灵敏度和争用性测量，以预测各种共址场景下的性能干扰。该方法能预测准确度也不错，同时也能够在实际运用中实现，但是也不适用于不同的NUMA系统，对于其他的系统，仍需要其他工作。

Baptiste等提出了AsymSched设计方案，一种新的线程和内存放置算法，其目标在于最大化通信线程的带宽。节点之间如何连接，连接的带宽不同，都会影响NUMA系统的性能，而随着NUMA系统节点的增加，这种影响更大。该方案虽然能够适用于不同的系统，并且容易执行，但是它没有预测性能准确度较差，此外，该方案也仅仅考虑了网络的非对称互连，没有考虑其他共享资源。

这篇论文提出的虚拟核到物理核的映射系统，弥补了上述方案或系统的不足。该系统考虑了多种共享资源如缓存，存储，带宽。根据机器学习，先实际测量两种映射配置的性能作为输入再预测，相比于比直接利用硬件性能事件，不仅性能预测准确率增高，而且更能适用于不同的NUMA系统，另外不需要大量的人力，也能实际应用。

二、数据中心节能技术

数据中心能耗成为其一大开销。如何降低数据中能耗也成为了研究的热点。目前主要的降低能耗思想分为两种，一是降低设备运行功率，如动态频率电压缩放（DVFS），即根据CPU需要的运行速率调整电压和频率；链路状态自适应（SLA），即改变链路的传输速率；双速磁盘，即在访问频率低时，降低磁盘运转速率。另一种是关闭空闲的设备，如动态电源管理（DPM），在CPU不用时关闭电源；网络流量合并（NTC），将流量合并到少数链路和交换机上，关闭不使用的链路和交换机；服务器负载合并（SLC），将工作负载集中到少数服务器以关闭其余的服务器。除去上述两种主要思想，还有数据中心节能方案，如将数据中心建立在寒冷地区，减少利用空调降温的能耗；利用可再生能源对数据中心进行部分供电，降低不可再生能源的消耗；改进数据中心的建筑构造，提高空调降温效率；保留热页面的虚拟机迁移降低迁移能耗等等。

随着虚拟机的出现，由于其具有隔离性，不依赖硬件，可迁移等优点，广泛应用于数据中心。因此大量的基于虚拟机迁移关闭未使用的服务器成为新热门的节能技术方向。Dabbagh等提出服务器的负载预测方法，根据历史负载变化与设定负阈值，进行虚拟机迁移，避免了由于过载导致的服务质量下降。KURDI等提出了基于蝗虫的生物启发式虚拟机迁移算法，其目标在于减少数据中心弱性能服务器的使用，因为较弱性能的服务器往往洗能效也偏低。虚拟机迁移能耗设计到虚拟机内存的大小和网络带宽，因此Wang等考虑了从CPU利用率，内存大小和网络带宽三个方面筛选虚拟机迁移。为减小宕机时间，需要采用迭代虚拟机内存迁移技术，在这种迁移过程中，也会影响虚拟机的性能，Liu等提出了EQVC算法，减少一些不必要的虚拟机迁移，不必要的定义为不进行虚拟机迁移带来性能损失小于进行虚拟机迁移的性能损失。另外还有不少研究人员针对服务质量、资源利用和能耗的多目标优化提出算法。

近些年，容器技术变得热门，由于其相对于虚拟机，更轻量化，更易于部署的优点，也在逐渐应用于数据中心。虚拟机迁移的不少技术和思想也可以应用于容器的迁移，这已经有不少的研究。

这篇论文是解决容器的虚拟核到物理核的映射问题，它考虑的是一定能耗条件下，可以达到的最大性能值，即最大化性能与能耗的比值，但是该配置只考虑的容器中不同虚拟核的信息交换，没有考虑容器之间的信息交换。而由于某个服务需要多个容器，多个容器会分布在同一个服务器或不同的服务器上，容器之间的交流也影响着服务的性能。也就是说，这篇论文只解决了容器的最大性能问题，没有解决服务的最大性能问题。现在已经有不少研究，以及之后研究方向，也就是多目标的优化问题，小的方面来说，如上述的容器性能最大化，和容器信息交流性能最大化。大的方面来说，不仅要考虑性能，也要考虑能耗，如容器的能耗，容器的交流，容器失效的重新部署。

**参考文献**

[1]Snavely A . Symbiotic Jobscheduling for a Simultaneous Multithreading Processor[J]. Acm Sigops Operating Systems Review, 2000, 34(5):234-244.

[2]Federova A , Seltzer M I , Smith M D . Improving Performance Isolation on Chip Multiprocessors via an Operating System Scheduler[C]// International Conference on Parallel Architecture & Compilation Techniques. IEEE, 2007.25-38.

[3] Merkel A , Stoess J , Bellosa F . Resource-conscious scheduling for energy efficiency on multicore processors[C]// European Conference on European Conference on Computer Systems. DBLP, 2010.153-66.

[4] Dashti M , Fedorova A , Funston J , et al. Traffic Management: A Holistic Approach to Memory Placement on NUMA Systems[J]. Acm Sigplan Notices, 2013, 41(1):381-394.

[5] Goodman D , Varisteas G , Harris T . Pandia : Comprehensive contention-sensitive thread placement[C]// Twelfth European Conference. 2017.254-269

[6] Lepers B , Vivien Quéma, Fedorova A . Thread and Memory Placement on NUMA Systems: Asymmetry Matters[C]// Usenix Conference on Usenix Technical Conference. USENIX Association, 2015.277-289

[7] Zhang Y , Laurenzano M A , Mars J , et al. SMiTe:Precise QoS Prediction on Real-System SMT Processors to Improve Utilization in Warehouse Scale Computers[J]. 2014.406-418

[8] Kurdi H A , Alismail S M , Hassan M M . LACE: A Locust-Inspired Scheduling Algorithm to Reduce Energy Consumption in Cloud Datacenters[J]. IEEE Access, 2018, 6.

[9] Liu Y , Sun X , Wei W , et al. Enhancing Energy-Efficient and QoS Dynamic Virtual Machine Consolidation Method in Cloud Environment[J]. IEEE Access, 2018, 6(99):1-1.

[10] Dabbagh M , Hamdaoui B , Guizani M , et al. An Energy-Efficient VM Prediction and Migration Framework for Overcommitted Clouds[J]. IEEE Transactions on Cloud Computing, 2016:1-1.