Scheduling using Optimization Decomposition in Wireless Network with Time Performance Analysis

Aparna.C[1], Kavitha.V.kakade[2]

M.E Student, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India [1]

Assistant Professor, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India [2]

ABSTRACT: Virtualization is a key technology underlying multi server computing platforms, where applications encapsulated within Virtual Machines are dynamically mapped onto a pool of physical servers. In this paper, we argue that multi server providers can significantly lower operational costs, and improve hosted application performance, by accounting for affinities and conflicts between co-placed Virtual Machines. The estimated Virtual Machine size is the basis for allocating resources commensurate with demand. In contrast to the traditional practice of estimating the size of Virtual Machines individually, we propose a joint-virtual machine provisioning approach in which multiple virtual machines are consolidated and provisioned together, based on an estimate of their aggregate capacity needs. This new approach exploits statistical multiplexing among the workload patterns of multiple virtual machines, i.e., the peaks and valleys in one workload pattern do not necessarily coincide with the others. Thus, the unused resources of a low utilized virtual machine can be borrowed by the other co-located virtual machines with high utilization. Compared to individual-virtual machine based provisioning; joint-virtual machine provisioning could lead to much higher resource utilization. This paper presents three design modules to enable such a concept in practice. Specifically, a performance constraint describing the capacity need of a virtual machine for achieving a certain level of application performance; an algorithm for estimating the aggregate size of multiplexed virtual machines; a virtual machine selection algorithm that seeks to find those virtual machine combinations with complementary workload patterns.

KEYWORDS: Virtual machines, Self organizing multi servers, Scheduling.

I. INTRODUCTION

Multi server platforms face two competing requirements from the perspectives of the multi server provider and the multi server user, respectively. The multi server provider would like to minimize power and cooling costs, which form a large portion of their operational costs. To achieve this goal, server consolidation can be used to minimize the number of physical servers required for hosting a set of applications. However, consolidation is often undesirable for multi server users, who are seeking maximum performance and reliability from their applications. Under provisioning of physical resources to applications may indirectly increase costs if frequent SLA violations result in lost business1. The key factor that determines the provisioning requirements of a Virtual Machine is its physical footprint: the amount of physical resources it consumes in terms of CPU time, storage, network bandwidth, and power. Existing consolidation and dynamic placement techniques have largely treated the physical footprint of a Virtual Machine as a location-independent measure. That is, it is generally assumed that the footprint of a Virtual Machine will be the same regardless of which physical machine it is placed on. This assumption is reasonable for a homogeneous environment, where physical machines are identical and most Virtual Machines are running the same OS and applications. However, in a multi server environment, we expect a diverse collection of applications to share a resource pool composed of heterogeneous resources. Applications will vary significantly in terms of resource requirements, access patterns and inter-dependencies; and the physical hosts will also vary in terms of their resource capacities and data affinities.

In such an environment, we contend that the physical footprint of a Virtual machine is highly malleable; that is, a virtual machine's physical resource consumption is heavily dependent on its location, the characteristics of its physical host, and its interactions with other virtual machines. A key factor for achieving economies of scale in a compute multi server is resource provisioning, which refers to allocating resources to virtual machines to match their workload. Typically, efficient provisioning is achieved by two operations: static resource provisioning. Virtual machines are created with specified size and then consolidated onto a set of physical servers. The virtual machine capacity does not change; and dynamic resource provisioning. Virtual machine capacity is dynamically adjusted to match workload fluctuations. Static provisioning often applies to the initial stage of capacity planning. It is usually conducted in offline and occurs on monthly or seasonal timescales. Such provisioning functionality has been included in much commercial multi server management software. In both static and dynamic provisioning, virtual machine sizing is perhaps the most vital step. Virtual machine sizing refers to the estimation of the amount of resources that should be allocated to a virtual machine. The objective of virtual machine sizing is to ensure that virtual machine capacity is commensurate with the workload. While over-provisioning wastes costly resources, under-provisioning degrades application performance and may lose customers.

Virtual machine sizing is done on a virtual machine-by-virtual machine basis, i.e., each virtual machine has an estimated size based on its workload pattern. In a significant departure from such an individual-virtual machine based approach, we advocate a joint-Virtual Machine provisioning approach in which multiple virtual machines are consolidated and provisioned based on an estimate of their aggregate capacity needs. Conceptually, joint-Virtual Machine provisioning exploits statistical multiplexing among the dynamic virtual machine demand characteristics, i.e., the peaks and valleys in one virtual machine's demand do not necessarily coincide with the other virtual machines. The unused resources of a low utilized virtual machine can then be directed to the other co-located Virtual Machines at their peak utilization. Thus, Virtual Machine multiplexing potentially leads to significant capacity saving compared to individual-Virtual Machine based provisioning. The savings achieved by multiplexing are realized by packing Virtual Machines more densely into hardware resources without sacrificing application performance commitment. While this increases the overall consolidation ratio, the additional virtualization overheads associated with scheduling somewhat higher number of Virtual Machines is generally minimal as long as the virtual machine footprints fit in the provisioned capacity. The savings with our joint-sizing approach are up to 40% according to our analysis on the utilization data from a production data center.

In this work, we address these questions in detail. Specifically, the primary contributions of this work are:

•   We introduce a Service-level-agreement (SLA) model that map application performance requirements to resource demand requirement. We propose a systematic method to estimate the total amount of capacity for provisioning multiplexed Virtual Machines. The estimated aggregate capacity ensures that the SLAs for individual Virtual Machines are still preserved.

•   We present a virtual machine selection algorithm that seeks to find those Virtual Machines with the most compatible demand patterns. The identified virtual machine combinations lead to high capacity savings if they are multiplexed and provisioned together.

•   We illustrate effective and feasible applications of the proposed technique for capacity planning and for providing resource guarantees via virtual machine reservations. Both applications can be easily employed in existing multi server and virtualization management infrastructures with minimal intrusion and substantial benefits in return.

II. RELATED WORK

In [4] Authors derive approximation ratiosto the sum rate optimal value by studying the solutions to two related problems, sum rate maximization using an SIR-approximation and SIR-max-min weighted SIR-optimization, thus obtaining a characterization ofthe power controlled rate region and its convexity property invarious asymptotic regimes.In [5] Authors proposed a general framework that allows the development of distributed mechanisms to achieve full utilization of multi-hop wireless networks for randomized routing, schedulingand flow control scheme that

is applicable to a large class of interference models.In [6] Authors presented two low-complexity distributed algorithmsfor scheduling in multi-hop wireless networks.

III. BENEFITS AND CHALLENGES

We envision benefits from both a multi server provider's perspective as well as a multi server user's perspective. From a multi server provider's perspective, reshaping the physical footprint to allow higher consolidation can help reduce hardware, energy and cooling costs. In addition, if such consolidation can be done dynamically without affecting the performance of applications, then maintaining user SLAs and achieving further cost savings in a time dependent manner can help substantially reduce provider costs and increase their profits. As an example, existing work in memory consolidation has shown that co-placing Virtual Machines with similar memory profiles can lead to substantial reduction in server memory usage. Note that this consolidation can be achieved in a manner that is completely transparent to applications hosted within the Virtual Machines.

The Virtual Machines running a distributed application can be reshaped to improve the application's performance, without requiring additional resources or support from the multi server provider, and then the user can achieve higher performance at no additional cost. To demonstrate the potential benefit of footprint reshaping for a multi server user, we ran a motivating experiment to show the impact of network footprint reduction on application performance. In this experiment, we used a micro benchmark to measure the time to send a number (1K-100K) of 100 KB files between two Virtual Machines hosted on different physical servers, and compared it to the time to transfer the same amount of data between Virtual Machines co-located on the same server. This experiment was conducted on a pair of identically-configured workstations running Xen 3.3.0, connected via 100 Mbps Fast Ethernet. As expected, our results show that as the quantity of data grows, the disparity between the two placements grows, with a reduction in transfer time of about 82. for 10K file transfers. Clearly, the benefit of co-placement in this case is highly dependent on the traffic volume and link bandwidth between the physical servers. In another experiment, we found that the make span of the Intel MPI benchmarks on a 1 Gbps network dropped from 646s, when the 3 Virtual Machines participating in the computation were distributed amongst 3 distinct servers, to 195s when the 3 virtual machines were co-located on the same physical server.

IV. METHODOLOGY

Virtual Machine multiplexing and joint-Virtual Machine provisioning approach is composed of three function modules, which collectively capture the necessary steps for defining the multiplexed Virtual Machines, and their individual and joint capacity requirements. We describe how these three modules cooperate within a general resource provisioning framework. The virtual machine selection algorithm identifies virtual machine combinations that achieve high capacity savings if provisioned together. The selection criterion is how complementary the virtual machine demand patterns are. Highly complementary virtual machines will be grouped together by the selection algorithm. Eventually the selection algorithm partitions virtual machines into multiple sets. Those virtual machines in the same set will be consolidated onto the same physical server and thus can be considered as a *super VIRTUAL MACHINE*. To provision such a super virtual machine, we first need to calculate its aggregate capacity need. We introduce a SLA model and a joint-Virtual Machine sizing algorithm.

The SLA model defines a relation between the virtual machine capacity and the performance level of the applications running on the virtual machine. The SLA model makes it convenient to derive a constraint on the super virtual machine capacity simply based on specified SLA for individual virtual machine. Based on the derived constraint and the aggregate workload of the super virtual machine, the joint-Virtual Machine sizing algorithm proceeds to calculate the super Virtual Machine's capacity need, which is the minimum amount of resources that should be allocated to the super virtual machine without degrading individual virtual machine's SLA. We apply such a virtual machine multiplexing approach and demonstrate its benefits in two multi server management operations. The first application is virtual machine consolidation. We identify compatible virtual machines, provision them jointly in a compute multi server and significantly reduce hardware requirement. Second, we define a *joint reservation* model to provide virtual machine-level resource guarantees in a virtualized environment. By identifying compatible virtual machines and their SLA, we are able to derive a joint reservation level based on individual virtual machine

reservations. We group compatible virtual machines in resource pools, and enforce joint reservations at the resource pool level. All of these enable dramatically improved virtual machine consolidation ratio in the multi server.

A. *Virtual Machine consolidation*:

Virtual machine consolidation is performed when a virtual machine controller needs to create and deploy a set of virtual machines on a set of physical hosts. The goal of virtual machine consolidation is to determine a mapping of virtual machines to physical hosts such that the minimum number of hosts is used. Existing virtual machine consolidation schemes consist of two steps: estimating the future size for each virtual machine, and placing virtual machines on physical hosts. The first step, estimating virtual machine size, is usually solved by first forecasting the future workload, then finding a capacity size that can sufficiently cover the forecasted workload. The second step, virtual machine placement, usually requires solving a bin packing type of problem. Specifically, since each virtual machine carries a size and each physical host has fixed capacity, the VIRTUAL MACHINE placement problem is equivalent to packing items into the smallest number of bins without violating the size limit on each bin. Virtual machine placement is tackled by either heuristics or solving an integer programming problem. By exploiting virtual machine multiplexing, it is possible to achieve even more compact consolidation. The only necessary change is to replace the first step in the above procedure with the proposed three building blocks. Briefly speaking, the virtual machine controller first applies the proposed SLA model to describe the performance requirement for each virtual machine. It then runs the virtual machine selection algorithm to partition virtual machines into virtual machine groups. The joint-Virtual Machine sizing algorithm is employed to determine the capacity being allocated.

B. *Reshaping for System-wide Goals*:

A control system for multi server platforms can utilize these principles to achieve higher measures of system-wide objectives such as power-savings and performance, through affinity-aware intelligent placement and migration. Given a set of virtual footprints, identifying an "optimal" placement across multiple resource dimensions is an instance of the bin packing problem, which is known to be NP-hard. Many placement and migration algorithms have been developed to solve this problem in the context of load-balancing. These algorithms would have to be tailored to incorporate the footprint reshaping principles described above, for instance, by including affinity/conflict information as additional constraints to the optimization problem. Alternatively, these principles can guide heuristics to enhance the initial solutions obtained from these algorithms. While existing algorithms coupled with the principles described above provide guidance for reducing the physical footprint along individual resource dimensions, reshaping the virtual machine along one resource dimension can have adverse consequences along other resource dimensions.

The optimal placement is dependent on higher-level system policies, which specify requirements in terms of potentially conflicting goals such as power-savings, performance and reliability. For example, a public multi server provider may desire to minimize hosted virtual machines' consumption of power, bandwidth and memory, even if it comes at the expense of increased CPU time and disk space. A private multi server, featuring an abundance of hardware and fast interconnects, might seek to minimize power consumption and execution time, regardless of bandwidth, memory and disk utilization. To address these challenges, the footprint shaping algorithms will need to incorporate optimization techniques which account for inter-resource trade-offs and conflicting objectives. Finally, in a large-scale system, control decisions would likely need to be performed in a decentralized manner.

V. CONCLUSION

This work advocates leveraging virtual machine multiplexing to improve resource utilization in compute multi servers. The benefit of virtual machine multiplexing is that when the peaks and troughs in multiple virtual machines are temporally unaligned, these virtual machines can be consolidated and provisioned together to save capacity. This paper presents three design modules that enable the concept in practice. Specifically, a new SLA model reflects application performance requirements; a joint-Virtual Machine sizing technique that estimates the aggregate capacity needs for multiplexed virtual machines; and a virtual machine selection algorithm for identifying most compatible virtual machine combinations. The proposed design modules can be seamlessly plugged into existing resource provisioning applications. The Experiments based on data from an operational multi server demonstrate that the proposed joint-Virtual Machine provisioning significantly outperforms traditional approaches. We showed how multi server providers

and users can benefit from increased consolidation and application performance by dynamically reshaping Virtual Machines' physical resource consumption through intelligent placement and migration. We described how to identify opportunities for reshaping by characterizing virtual machines in terms of a location-independent virtual footprint, and then presented three general principles for minimizing a Virtual Machine's physical footprint with respect to memory, network, and disk and power consumption. Finally, we discussed challenges associated with applying these principles in practice in a multi server environment.

REFERENCES

1. M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," Proc. IEEE, vol. 95, no. 1, pp. 255–312, Jan. 2007.
2. C. W. Tan, D. Palomar, and M. Chiang, "Exploiting hidden convexity for flexible and robust resource allocation in cellular networks," in Proc. IEEE INFOCOM, 2007, pp. 964–972.
3. C. W. Tan, D. Palomar, andM. Chiang, "Energy-robustness tradeoff in cellular network power control," IEEE/ACM Trans. Netw., vol. 17, no. 3, pp. 912–925, Jun. 2009.
4. C. W. Tan, M. Chiang, and R. Srikant, "Fast algorithms and performance bounds for sum rate maximization in wireless networks," in Proc. IEEE INFOCOM, 2009, pp. 1350–1358.
5. A. Eryilmaz, O. Asuman, and E. Modiano, "Polynomial complexity algorithms for full utilization of multi-hop wireless networks," in Proc. IEEE INFOCOM, 2007, pp. 499–507.
6. A. Gupta, X. Lin, and R. Srikant, "Low-complexity distributed scheduling algorithms for wireless networks," IEEE/ACM Trans. Netw., vol. 17, no. 6, pp. 1846–1859, Dec. 2009.
7. C. W. Tan, S. Friedland, and S. H. Low, "Spectrum management in multiuser cognitive wireless networks: Optimality and algorithm," IEEE J. Sel. Areas Commun., vol. 29, no. 2, pp. 421–430, Feb. 2011.