Characterization of dynamic resource consumption for interference-aware consolidation

Markus Hähnel

Chair for Computer Networks, Faculty of Computer Science
Technical University of Dresden, 01062 Dresden, Germany
E-Mail: markus.haehnel1@tu-dresden.de

Abstract. Nowadays, our daily live concerns the usage of Information Technology, increasingly. As a result, a huge amount of data has to be processed which is outsourced from local devices to data centers. Due to fluctuating demands these are not fully utilized all the time and consume a significant amount of energy while idling. A common approach to avoid unnecessary idle times is to consolidate running services on a subset of machines and switch off the remaining ones. Unfortunately, the services on a single machine interfere with each other due to the competition for shared resources such as caches after the consolidation, which leads to a degradation of performance. Hence, data centers have to trade off between reducing the energy consumption and certain performance criteria defined in the Service Level Agreement. In order to make the trade off in advance, it is necessary to characterize services and quantify the impact to each other after a potential consolidation. Our approach is to use random variables for characterization, which includes the fluctuations of the resource consumptions. Furthermore, we would like to model the interference of services to provide a probability of exceeding a certain performance criterion.

Keywords: dynamic workload, characterization, resource consumption, consolidation, interference, energy-efficient computing, HAEC

1 Introduction

In their daily lives, citizen all over the globe increasingly use cloud-based information and communication technology services. As a result a huge amount of data has to be processed by an ever-growing multitude of power consuming servers in data centers. At the same time, using data centers is an effective approach to combine all the competence of building, managing, maintaining, etc. the infrastructure. Nevertheless, the data center has to provide a certain quality to his costumers which are paying for the service. In the context of Information Technology (IT) these demands are specified by a Service Level Agreement (SLA) between the costumer and the data center [1]. Commonly, the SLA contains limits of some performance criteria such as latency or response time of the hosted applications.

However, data centers consume a significant amount of energy unproportional to the workload [14]. Therefore, in order to improve the energy efficiency (work done per energy) in the data centers, different approaches have to be taken. The addressed points divide mainly into two aspects, namely IT itself and the periphery. First, the periphery like cooling must be minimized. Modern examples of Facebook and Google show that it is possible to reduce the proportion for the periphery to less than 10 % of the data center's overall power consumption [2]. Second, the way the machines are utilized has to be optimized for best energy efficiency. The overall power consumption should be proportional to the workload but servers consume up to 50 % of their peak power when they are idling. Hence, the best energy efficiency is reached only when the running servers are all fully utilized. Nevertheless, a data center has to perform accordingly to the current demands even if the amount of workload fluctuates over the time (see Figure 1 [6]). To satisfy the rare high demands, it is designed to accomplish a certain degree of peak performance. However, the average utilization is much lower, for instance, only 10 % for a university data center or 25 % for Wikipedia [14]. Furthermore, when a service is running, it does not fully utilize a system. An approach for avoiding underutilized servers is to consolidate the currently running services on a subset of machines [3], [4], [11], [15]. Thanks to several live-migration-techniques with very low downtimes the migration can be suitable even for very high SLAs [10]. Afterwards, this subset of machines operates more efficiently because of a higher utilization. Subsequently, the remaining systems which are idling can be switched off. The energy savings exceed the migration costs after only a short period [10].

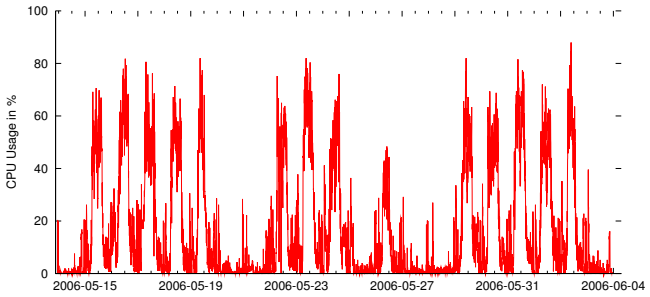

Fig. 1. Fluctuations of the workload (represented by the example of the overall CPU usage) over the time [6]. Variation between day (high workload) and night (low workload), for instance, are addressed by workload consolidation to a subset of servers.

After consolidation, the services will be affected by each other. For instances, even when a service was assigned exclusively to one core of a processor, it will be influenced by the services running on the other cores due to the limited size of shared last level cache (LLC). Obviously, the contention of different services, called interference, depends on several parameters because CPU, memory, disk,

and network do not operate independently from each other. For example, all reads and writes to the disk are cached in the memory. Unfortunately, unlike the CPU which can be utilized from 0 to 100 %, there is no common measure to quantify the activity to memory for instance. If we want to consolidate services with different resource demands to maximize the entire utilization of the target system, we have to answer the following two questions.

Research Question 1: *Which parameters characterize a service in terms of its resource consumptions?*

Furthermore, the service consolidation represents a optimization problem. On the one hand, the data centers would like to power on only as few as possible servers to save energy. On the other hand, the service consolidation introduces interference, and hence affects the SLAs. Due to the time required for rebooting additional servers and re-migrate a service, the data center has to trade off between energy savings and performance in advance. Therefore, it is necessary to estimate the influence and resulting performance decrease. This leads to Research Question 2.

Research Question 2: *What is the impact on the performance of one service to another after consolidation?*

We are going to address the characterization as well as the estimation of interference with random variables. This enables to describe the dynamic of services while the services are consolidated. Furthermore, the mathematical formalism for describing random variables can yield criteria for the SLA. The rest of the paper is organized as follow: In section 2 we summarize a few exemplary publications with related work. Afterwards we discuss our approach and methods in section 3. Finally, we conclude our main concerns in section 4.

2 Related work

Several studies address the consolidation of Virtual Machines (VMs), particularly, resource contention between co-located VMs, and the performance degradation to keep performing a certain SLA [7]–[9], [12], [13]. To minimize the computation effort at runtime, Govindan *et al.* [7] identify some VMs with characteristic interference signature and measure their mutual influence. Afterwards, every productive VM is mapped to one of these characteristic VMs so that the performance after consolidation can be predicted. This strategy needs a low computation effort at runtime, but requires to define a set of characteristic VMs and discard details of productive VMs by mapping them to this limited set.

Another approach is done by Srinivasan and Bellur [12]. They model immediately the job completion time. Independent of the concrete combination of consolidated VMs they rely on a general parameter, namely the CPU utilization. It is split into an independent and dependent part because the duration of a task does not depend only on the CPU utilization. An advantage of this approach is the possibility to model the power consumption based on the CPU utilization. Additionally, they include the current frequency of the processor for their model.

Roytman *et al.* [9] observe that the CPU independent part of performance degradation comes mainly from the contention in shared caches and memory band widths. Again, all active VMs are mapped to a finite number of classes to predict the degradation. Afterwards they introduce a metric for the costs of consolidation: Degradation of VMs divided by the number of VMs. In the performance mode (minimize resource costs under performance constraints), they calculate this metric for all combinations, order them in descending order, and locate the VM sets to the servers. Considering all combinations include the optimal solution, but due to the mapping to a predefined characteristic VMs this optimal solution is only approximated, again. The performance mode yield an energy saving of 30 % by switching of the remaining unused servers. Alternatively, they introduce an eco mode (minimize degradation under cost constraints). VMs are iteratively permuted until no improvement could be achieved or a limit of permutations is reached.

Another characterization is used by Verboven *et al.* [13]. They classify each workload by CPU utilization, cache hit/miss-rate, and disk-I/O. Unfortunately, network-I/O is omitted and only one vCPU per VM is used as common assumption in literature. First, they measured each VM running alone and afterwards co-locate evermore VMs. Therefore, they describe the performance lost of consolidation. Finally, they propose a scheduling algorithm as consequence. The characterization by generic parameters belonging to the VM makes the mapping to a finite predefined set of VMs obsolete. On the other hand, the effect of a VM on another VM has to be separately considered for each of them.

As discussed above, interference depends strongly on the cache and the memory. Hence, Kim *et al.* [8] focus their investigation on the LLC and the memory bus. They observe that the runtime consists of calculation time and the access times of the cache levels and memory. The latter one can be estimated by the LLC misses. Both the number of LLC-misses and the memory access time increases when two services were consolidated. Thus, Kim *et al.* [8] define the interference intensity and the interference sensitivity. The intensity is described by the LLC-misses, LLC-references, and the execution duration. It is a measure of how strong the service impacts another service. The sensitivity is described by the ratio of number of cycles waited for memory and overall cycles. It is a measure of how strong the service is impacted by another one. Finally, they co-locate services having a high interference intensity and services having a low interference sensitivity.

All of these approaches consider only static parameters and static states of services. Dynamic workload, and hence varying utilizations of hardware is neglected. Therefore, SLAs can be violated while peaks of high workload and energy can be wasted while weak utilization.

3 Methods

We wish to minimize the power consumption of a data center by consolidating services to make the power consumption proportional to the workload [5]. But

consolidation creates contention between services, which in term may degrade performance and increase power consumption. Therefore, we wish to develop a strategy to measure contention, called interference. We propose a stochastic model to identify and model interference. We employ performance indicator parameters such as CPU utilization, cache miss rate, cache hit rate, instructions retired, etc. for our stochastic model. The model cover the dynamic of a service. For example, most of the services do not have static workload. A web service becomes active only when a user initiates a request. Also, the degree of resource utilization varies; sometimes just a few users access the web service (e.g. a homepage) and sometimes many users put requests to the service. In this example, the requests are a random workload. In Figure 1 is shown that it is possible to predict the workload changes within a time window of one day (e.g. weekday and weekend). A bit more fine-grained prediction can also be done for night and day. In contrast to this, it is not possible to deduce the upcoming workload based on the workload before for the time scale of one hour. As a result, all performance indicator parameters, such as CPU utilization, fluctuate. The idea is to describe such a changing value as random variable, denoted by X. There are several descriptive formalisms for random variables: expectation value, cumulative distribution function (CDF), probability density function (p.d.f.), etc. The expectation value $E\{X\}$ gives the average of the random variable. Of course, we are interested in a more detailed description which is given by the CDF $F_X(x)$. It gives the probability P that the random variable is lower or equal to a certain value x: $F_X(x) = P\{X \leq x\}$. The derivation of the CDF is called p.d.f. $f_X(x)$. The p.d.f. gives the probability $f_X(x)\,\mathrm{d}x$ that X is in the interval $[x, x + \mathrm{d}x]$. Because of the probability of 1 that X is anywhere, the p.d.f. is normalized: $\int f_X(x)\,\mathrm{d}x = 1$.

As mentioned in section 1 it is not always possible to describe the resource utilization only by a single parameter. Especially, the memory utilization rely on the number of different events like retired instructions and LLC misses. The stochastic formalism enables us to combine such performance indicator parameters described by random variables, including their dynamic expressed by the p.d.f. Furthermore, we can estimates the degree of contention after consolidation. As a simple example, we consider the CPU utilizations X and Y of two independent services (see Figure 2). The statistics of the random variables X and Y can be obtained by data mining before the consolidation. For running example, we approximate the usual complex p.d.f. by a normal distribution $f_X(x) = N(\mu_X = 45\,\%, \sigma_X = 15\,\%)$ and $f_Y(y) = N(\mu_Y = 30\,\%, \sigma_Y = 10\,\%)$, respectively. After consolidation of both services on a server which was idling before, we would expect an overall CPU utilization of $Z = X + Y$ with the p.d.f.

$$f_Z(z) = \frac{1}{\xi} \int f_X(x) \cdot f_Y(z - x)\,\mathrm{d}x = N(\mu_X + \mu_Y, \sigma_X + \sigma_Y)\,.$$

The ξ is just a normalization factor to preserve the normalization of the obtained p.d.f. After integration of $f_Z(z)$, we obtain the CDF $F_Z(z)$ which yields the probability $1 - F_Z(100\,\%) = 15.6\,\%$ that the server is overloaded after the consolidation. In this case, the CDF can be used as a SLA.

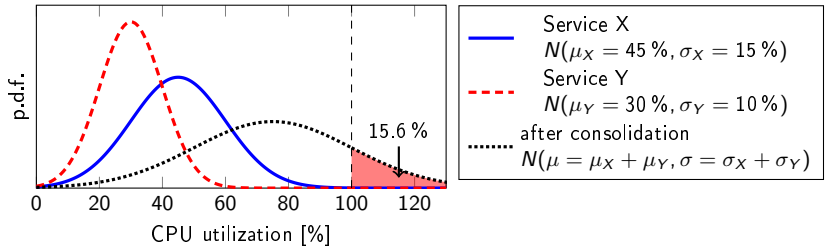

Fig. 2. Distribution functions of CPU utilization of Service X and Service Y before consolidation, and the expected overall CPU utilization after consolidation.

In the first step, we identify suitable parameters and patterns of their dynamic to characterize services regarding to their resource consumption. As second step, we investigate the dependencies and contentions of different resources. Finally, we try to estimate the performance decrease of services with resource consumptions characterized as complementary based on the interferences that we have found.

Our approach will serve as new metric for the optimization problem between saving energy by service consolidation and introduced interference. It defines a new optimal solution which has to be compared to already existing consolidation strategies and their implementations.

4 Conclusion

The significant contribution of data centers to the world energy consumption makes it necessary to improve their efficiency as much as possible. A common strategy is to scale the number of running systems to the current demands by consolidating the active services on a subset of servers and switching off entirely the remaining idling systems. However, even these machines are running efficiently for themselves only when they are fully utilized. This includes a minimization of idle times as well as utilizing other resources than the CPU such as memory, disk, and network. Current models of characterization assume only static aspects of resource consumption. We aim to find an approach which also includes the dynamic of a service by describing varying values as random variables.

Further, we wish to model the contention of different resources to estimate, based on our service characterization, the interference after consolidation. Thus, it will be possible to consolidate services with respect to both full utilization of all resources of the server system for best efficiency, and SLAs.

Acknowledgement. This work is supported by the German Research Foundation (DFG) within the Collaborative Research Center SFB 912 – HAEC. Special thanks to my supervisor Dr. Waltenegus Dargie and my colleague Frehiwot Melak Arega for their constructive feedback and inspiring discussions.

References

1. T. G. Berger, *Service Level Agreements*. VDM, 2007.
2. A. S. Brown, "Keep it cool! inside the world's most efficient data center", in *The Bent of Tau Beta Pi*, 2014.
3. G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services", in *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2008.
4. E. Elnozahy, M. Kistler, and R. Rajamony, "Energy-efficient server clusters", in *Power-Aware Computer Systems*, Springer Verlag, 2003.
5. G. Fettweis, W. E. Nagel, and W. Lehner, "Pathways to servers of the future: highly adaptive energy efficient computing (haec)", in *Conference on Design, Automation and Test in Europe (DATE'12)*, 2012.
6. D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper, "Workload analysis and demand prediction of enterprise data center applications", *IEEE International Symposium on Workload Characterization (IISWC)*, 2007.
7. S. Govindan, J. Liu, A. Kansal, and A. Sivasubramaniam, "Cuanta : quantifying effects of shared on-chip resource interference for consolidated virtual machines", in *2nd ACM Symposium on Cloud Computing*, 2011.
8. S. G. Kim, H. Eom, and H. Y. Yeom, "Virtual machine consolidation based on interference modeling", *The Journal of Supercomputing*, vol. 66, no. 3, 2013.
9. A. Roytman, A. Kansal, S. Govindan, J. Liu, and S. Nath, "Pacman: performance aware virtual machine consolidation", in *10th International Conference on Autonomic Computing (ICAC'13)*, 2013.
10. K. Rybina, W. Dargie, S. Umashankar, and A. Schill, "Modelling the live migration time of virtual machines", in *On the Move to Meaningful Internet Systems: OTM 2015 Workshops*, Springer International Publishing Switzerland 2015, 2015.
11. S. Srikantaiah, A Kansal, and F Zhao, "Energy aware consolidation for cloud computing", in *Power aware computing and systems*, 2008.
12. S. P. Srinivasan and U. Bellur, "Watttime: novel system power model and completion time model for dvfs-enabled servers", in *IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, 2015.
13. S. Verboven, K. Vanmechelen, and J. Broeckhove, "Black box scheduling for resource intensive virtual machine workloads with interference models", *Future Generation Computer Systems*, vol. 29, no. 8, 2013.
14. W. Zhang, S. Rajasekaran, and T. Wood, "Big data in the background: maximizing productivity while minimizing virtual machine interference", in *Workshop on Architectures and Systems for Big Data*, 2013.
15. Q. Zhu, J. Zhu, and G. Agrawal, "Power-aware consolidation of scientific workflows in virtualized environments", in *ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, 2010.