Power Efficient Management In Data Centers Using Server Consolidation For Green Computing

Gnanasundaram.R
ME
Computer Science & Engineering
Adhiyamaan Collage of Engineering, India

S.Suresh
Associate Professor
Computer Science & Engineering
Adhiyamaan Collage of Engineering, India

Abstract

A Cloud computing has revolutionized the Information Technology industry by enabling elastic On-Demand provisioning of computing resources. Cloud data centers consume enormous amounts of electrical energy resulting in high operating costs and carbon-di-oxide emissions. The goal is to improve the utilization of computing resources and reduce energy consumption under workload independent quality of service constraints. Virtual Machine consolidation leverages fine-grained fluctuations in the application workloads and continuously reallocates Virtual Machines to minimize the number of active physical nodes. Server virtualization is a prominent approach to consolidate applications from multiple applications to one server. With a growing concern on the considerable energy consumed by data centers, research efforts are targeting toward green data centers with higher energy efficiency. The results obtained have proved that Server Consolidation has greatly reduced the energy consumption without affecting the performance of the system.

Keywords: Virtualisation, Energy Efficiency, Cloud Computing, Energy Consumption, Allocation of VM's

I. INTRODUCTION

Many large-scale IT services are relying on Cloud infrastructures to host applications and to process data. Cloud computing allows computing resources to be provided as utilities, users can request CPU power, storage space, or access to applications and only pay for their use as needed. Resources which are no longer needed can be released at any time. From the user's perspective, Cloud computing allows access to resourceson demand, without the need to acquire, provision or maintain them. Furthermore, users only pay for what they actually use, and thus can make optimal utilization of resources .A Cloud service is physically hosted inside big datacenters, containing a large number of computing nodes. The energy requirement of the whole datacenter is a significant fraction of the total operating costs . Thus, reducing the energy consumption is becoming an important issue both for economical reasons (reducing costs) but also for making IT services environmentally sustainable. In this paper we address the problem of reducing the power consumption of Cloud infrastructures by moving VMs on a limited subset of the available (physical) computing resources, so that the remaining (idle) computing nodes can be switched to low power consumption modes. The process of aggregating services running on multiple servers into a reduced number of more powerful servers is known as server consolidation. In this paper we use the term VM consolidation to denote the consolidation of multiple VMs on a reduced number of physical hosts.

In recent years, virtualization has been changing the way information technology infrastructure in enterprise data centers is built. The need for large data centers arose due to demand for computational power . This computational power goes to running services. Cloud infrastructures like Amazon Elastic Compute Cloud and Microsoft Online Services provide resources for various computing needs.

Google and many others offer software as a service directly to a user's web browser . Shopping has been shifting more and more online . All these services require large amounts of servers and flexibility to satisfy the demands of an ever-growing user base. This growth has had a side effect. Lately, energy expenses in the data centers have been rising to the extent of possibly surpassing the actual hardware costs . All this calls for more energy efficient computing. In many cases of real life data servers, however, energy efficiency measures are conducted only when the infrastructure has already reached its maximum capacity. Even then, the focus of optimization has mostly been in hardware and infrastructure, not in operational methods, operating systems, or software . Hardware optimization is only part of the solution and software also must be taken into account when pursuing energy efficiency. Servers require new energy efficiency innovations in addition to the energy efficient hardware. Power consumption reduction in circuit techniques and hardware features among others, and also recognized the importance of software in energy efficiency. Virtualization is one solution to the problems. Using virtualization enables one to cut costs and enhance energy efficiency. There have been a lot of synthetic performance tests using virtualized systems.

Recently, with the rapid development of virtualization technology, such as VMware , Xen , KVM , OpenVZ, more and more data centers use this technology to build new generation data center architecture to supportcloud computing due to the benefits such as improving resource utilization, reducing costs, easing server management. What's more, server consolidation and live migration of virtual machine are two crucial methods to achieve load balancing and energy saving. Server consolidation which

allowing multiple servers running in a single physical server simultaneously is a main approach to achieve better energy efficiency of data center. It is because in doing so, server consolidation allows more physical servers to be turned off via live migrating the virtual machines to other unsaturated physical servers. Although virtual machine technology can improve the energy efficiency in data center, the overheads caused by virtualization and the efficiency of consolidation and migration strategies need to be investigated.

Cloud computing or simply a cloud is the realization of the paradigm of shifting the location of computing infrastructure to the network,with the aim of reducing both software and hardware resource management costs. Usage of clouds can split into three scenarios : Infrastructure as a Service (IaaS), Platform as a Service(PaaS) or Software as a Service (SaaS) . In IaaS scenario,virtualization is used as the customer wishes. At this point it is important to note that the foundation of cloud computing is formed by virtualization, as it is the technology which provides the capability to handle the resources as stated above .

As a generic term, virtualization means creating a virtual version of something, a simulation of the real version . In computing, it is "a technique for hiding the physical characteristics of computing resources from the way in which other systems, applications, or end users interact with those resources." Virtualization technologies are widely used in a variety of areas such as multiple operating system support, server consolidation, transferring systems from one computer to another, secure computing platforms and operating system development.

Basically, when something has been virtualized it is no longer bound to the physical realization. This in turn enables N:N relationships between applications and hardware: one is able to run multiple isolated applications on a single shared resource, but also a single application on multiple physical resources . One of these uses of multiple physical resources is live migration, which nowadays is one of the most important uses of virtualization techniques. Another use of virtualization is in application deployment and management.

Server consolidation by virtualization means running multiple applications in separate virtual containers hosted on single hardware. These virtual containers can be full virtual machines and the applications can be operating systems: in essence, one is able to consolidate multiple virtual machines into one real machine. In enterprise data centers, this has become an integral part of information technology planning to more efficiently utilize hardware and to reduce costs.

In this work, we first present a virtual machine based energy-efficient data center architecture for cloud computing and discuss the details. Then, we evaluate theperformance overheads of server consolidation and investigatethe consolidation efficiency which will affect the energy efficiency with varying degrees.

II. RELATED WORK

One of the first works, in which power management has been applied in the context of virtualized data centers, has been done by Nathuji and Schwan[1]. The authors have proposed an architecture of a data center's resource management system where resource management is divided into local and global policies. At the local level the system leverages the guest OS's power management strategies. The global manager gets the information on the current resource allocation from the local managers and applies its policy to decide whether the VM placement needs to be adapted. However, the authors have not proposed a specific policy for automatic resource management at the global level.

Kusic et al.[2] have defined the problem of power management in virtualized heterogeneous environments as a sequential optimization and addressed it using Limited Look ahead Control (LLC). The objective is to maximize the resource provider's profit by minimizing both power consumption and SLA violation. Kalman filter is applied to estimate the number of future requests to predict the future state of the system. However, the proposed model requires simulation-based learning for the application-specific adjustments, which cannot be implemented by Infrastructure as a Service (IaaS) Cloud providers, such as Amazon EC2. Moreover, due to the model complexity the execution time is not suitable for large-scale real-world systems. On the contrary, which does not require simulation based learning prior to the application deployment and allows the achievement of high performance even for a large scale as shown by experiments.

Srikantaiah et al. [3] have studied the problem of request scheduling for multi-tier web applications in virtualized heterogeneous systems to minimize energy consumption, while meeting performance requirements. The authors have investigated the effect of performance degradation due to high utilization of different resources when the workload is consolidated. They have found that the energy consumption per transaction results in a "U"-shaped curve, and it is possible to determine the optimal utilization point. To handle the optimization over multiple resources, the authors have proposed a heuristic for the multidimensional bin packing problem as an algorithm for the workload consolidation. However, the proposed approach is workload type and application dependent, whereas our algorithms are independent of the workload type, and thus are suitable for a generic Cloud environment.

Verma et al. [4] have formulated the problem of power-aware dynamic placement of applications in virtualized heterogeneous systems as continuous optimization: at each time frame, the placement of VMs is optimized to minimize power consumption and maximize performance. Like in [3] the authors have applied a heuristic for the bin packing problem with variable bin sizes and costs. Similarly to [1] live migration of VMs is used to achieve a new placement at each time frame. The proposed algorithms, on the contrary to our approach, do not support SLAs: the performance of applications can be degraded due to the workload variability. In their more recent work [5], Verma have proposed dividing VM consolidation strategies into static (monthly, yearly), semistatic (days, weeks) and dynamic (minutes, hours) consolidation. In the paper, the authors have focused on static

and semi-static consolidation techniques, as these types of consolidation are easier to implement in an enterprise environment. In contrast, in this work we investigate the problem of dynamic consolidation to take advantage of fine-grained optimization.

Gandhi et al. [6] have investigated the problem of allocating an available power budget among servers in a virtualized heterogeneous server farm, while minimizing the mean response time. To investigate the effect of different factors on the mean response time, a queuing theoretic model has been introduced, which allows the prediction of the mean response time as a function of the power-to-frequency relationship, arrival rate, peak power budget, etc.

In recent years the commercial, organizational and political landscape has changed fundamentally for data centre operators due to a confluence of apparently incompatible demands and constraints. The energy use and environmental impact of data centers has recently become a significant issue for both operators and policy makers. Global warming forecasts that rising temperatures, melting ice and population dislocations due to the accumulation of greenhouse gases in our atmosphere from use of carbon-based energy. Unfortunately, data centers represent a relatively easy target due to the very high density of energy consumption and ease of measurement in comparison to other, possibly more significant areas of IT energy use. Policy makers have identified IT and specifically data centre energy use as one of the fastest rising sectors. At the same time the commodity price of energy has risen faster than many expectations. This rapid rise in energy cost has substantially impacted the business models for many data centers. Energy security and availability is also becoming an issue for data centre operators as the combined pressures of fossil fuel availability, generation and distribution infrastructure capacity and environmental energy policy make prediction of energy availability and cost difficult .As corporations look to become more energy efficient, they are examining their operations more closely.

III. PROBLEM STATEMENT

As some of the developing countries are facing huge energy crisis, the energy consumed by data centers have a great effect on their overall production of energy. This huge consumption of energy by data centers not only shortens the supply of power energy to other businesses but also contributes towards the shortage for data centers themselves. This energy consumption also contributes towards waste of energy and environmental stewardship. There is a need to design a strategy that provides a solution to decrease the continuous demand and consumption of energy by data centers. This work proposes a new technique that combines the workload of multiple servers onto fewer servers by properly utilizing their efficiencies i.e. hardware and software efficiencies.

IV. PROBLEMS OF HIGH POWER AND ENERGY CONSUMPTION

Energy consumption by computing facilities raises various monetary, environmental ,and system performance concerns. A recent study on power consumption of server farms showed that in 2005 the electricity use by servers worldwide – including their associated cooling and auxiliary equipment – cost 7.2 billion dollars. The study also indicates that the electricity consumption in that year had doubled compared to the consumption in 2000. Clearly, there are environmental issues with the generation of electricity. The scope of energy efficient design is not limited to main computing components (e.g., processors, storage devices, and visualization facilities), but can expand into a much larger range of resources associated with computing facilities including auxiliary equipment, water used for cooling, and even the floor space occupied by the resources. While recent advances in hardware technologies including low-power processors, solid state drives, and energy-efficient monitors have alleviated the energy consumption issue to a certain degree, a series of software approaches have significantly contributed to the improvement of energy efficiency. These two approaches (hardware and software) should be seen as complementary rather than competitive. User awareness is another non-negligible factor that should be taken into account when discussing Green IT. This, in turn, has a direct relationship with energy consumption of not only core computing resources but also auxiliary equipment, such as cooling air conditioning systems. For example, a computer program developed without paying much attention to its energy efficiency may lead to excessive energy consumption and may contribute to higher heat emission resulting in increases in the energy consumed for cooling.

V. CONCLUSION

With the growing use of internet and requirement of data-storage and processing, the size of modern data centers has greatly increased. This has led to significant increase in the power consumption levels of the data centers. Moreover, the power consumption of data centers is approaching the limit imposed by thermal limitations of cooling solutions and power delivery. Also, since data centers are already consuming tens of Mega Watts, they are also stressing the capabilities of power generation facilities. As the complexity of operation of data centers increases, power management techniques which also ensure high-performance and low-costs are expected to become a crucial part of future enterprise architectures. We reviewed several techniques which have been proposed for reducing power consumption of data centers and classified them based on their characteristics

REFERENCES

[1] Nathuji R, Schwan K. Virtual power: Coordinated power management in virtualized enterprise systems. ACM SIGOPS Operating Systems Review 2007; 41(6):265–278.
[2] Kusic D, Kephart JO, Hanson JE, Kandasamy N, Jiang G. Power and performance management of virtualized computing environments via lookahead control. Cluster Computing 2009; 12(1):1–15.
[3] Srikantaiah S, Kansal A, Zhao F. Energy aware consolidation for cloud computing. Cluster Computing 2009; 12:1–15.
[4] Verma A, Ahuja P, Neogi A. pMapper: Power and migration cost aware application placement in virtualized systems. Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware (Middleware 2008), Springer, Leuven, Belgium, 2008; 243–264.
[5] Verma A, Dasgupta G, Nayak TK, De P, Kothari R. Server workload analysis for power minimization using consolidation. Proceedings of the 2009 USENIX Annual Technical Conference, San Diego, CA, USA, 2009; 28–28.
[6] Gandhi A, Harchol-Balter M, Das R, Lefurgy C. Optimal power allocation in server farms. Proceedings of the 11th International Joint Conference on Measurement and Modeling of Computer Systems, ACM New York, NY, USA, 2009; 157–168.