ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 3, March 2015

Power Optimization in Cloud Computing using different Methodologies

Nikhil Save[1], Prof. Varshapriya J.N.[2]

Student, Computer Engineering Department, VJTI, Mumbai, India [1]

Professor, Computer Engineering Department, VJTI, Mumbai, India [2]

Abstract: Cloud computing is the current technology used for sharing and accessing resources using web services or internet. It provides a scalable and cost effective environment. Large number of servers in the datacentres leads to huge consumption of power in the cloud computing scenario. Energy-related costs have become one of the major economic factors in IT datacentres, and companies and the research community are currently working on new efficient power aware resource management strategies. Optimization of power consumption is a key challenge for effectively operating a datacentre.

Keywords: Optimization, Machine Learning, DVFS, Power Saving.

1 INTRODUCTION

CLOUD computing is an on demand service in which shared resources, information, software and other hardware devices are provided according to the clients requirement at specific time. It's a term which is generally used in case of Internet. The whole Internet can be viewed as a cloud. Capital and operational costs can be cut using cloud computing. The Cloud computing model takes advantage of virtualization using computing resources allowing customers to use resources on-demand on a pay-as-you-go basis. Instead purchasing the costly IT infrastructure and dealing with the maintenance and up gradation of software and hardware, organizations can reduce operational cost by outsourcing their computational needs to the Cloud. The proliferation of Cloud computing has led in the establishment of large scale data centres containing huge numbers of computing nodes and consuming enormous amounts of electrical energy. [1] Before applying power optimization techniques we must be able to calculate the amount of power consumed by various components of the cloud environment. Different components which consume power are viz. CPU, memory and hard disks. Once the power consumption by different components is calculated we can optimize power by using Neural Network for load prediction, which predicts the upcoming load based on the past historical data. The servers can be monitored and given workload based on their reliability record and this data is used as a criterion while performing load balancing.

1.1 Current Issue in Power Saving

The energy consumed in a datacentre (server hardware) is saved by two techniques: dynamic voltage/frequency scaling (DVFS) and shutting down servers when not in use. In dynamic voltage/frequency scaling the power will be saved by adjusting the operating clock to scale down the supply voltages, which uses adaptive algorithm.

Although DVFS approach reduces power consumption, it depends on the hardware components settings to perform scaling tasks. The shutting down of servers will conserve more power but turning resources off in a dynamic environment causes high overhead which leads to a strong performance degradation.

Before applying power optimization techniques we must be able to calculate the amount of power consumed by various components of the cloud. Different components which consume power are viz. CPU, memory and hard disks. Among them, the most extensively studied is the CPU, as it is the component with the largest impact on the overall energy usage of physical servers. As DVFS technique is widely used in today's environment it is not effective way of power saving. In this proposed solution we will mainly be concentrating on different algorithms for shutting down the less loaded severs by simply transferring the load to an-other running server.

2 METHODOLOGIES FOR ENERGY AWARE SCHEDULING

Here we propose different methodologies of autonomic energy aware scheduling that dynamically adapts to varying task types and workloads, and even to varying infrastructure. The main contribution is to combine different technologies such as virtualization and consolidation (to move tasks between hosts), mathematical programming (to create and solve models of tasks, hosts, and constraints), and machine learning and data mining (to build these models from examples of past behaviours) [4].

2.1 Machine Learning Methodology

A cloud is a technology that can be viewed as a set of jobs or tasks to be distributed along a set of resources, so the main decisions to make are what resources are given to what jobs, assuring that each job receives sufficient resources to be executed properly with respect to global goals and customer requirements included in SLA.

In order to make the management more efficient, here we employ methods from Machine Learning. This is a subfield of the data mining area in charge of modelling systems from real examples of their historical behaviour.

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 3, March 2015

These models can then be used to predict future behaviours. The advantage of Machine Learning, as opposed to explicit expert modelling, is that it applies when systems are complex enough that no individual human expert can explore all relevant model possibilities, or in domains when no experts exist, or when the system is ever changing over time that models have to be rebuilt autonomously with time.

2.2 Consolidation Methodology

The relation between resource usage and power grows no proportionally and sub linearly. This explains the potential for power saving by consolidation. E.g. in a Intel Xeon 4-CPU machine, the power consumption (Watts/hour) when in all CPUs are IDLE is 235, when only 1 CPU is active is 267.8, and when 2, 3, and 4 CPUs are active, the power consumption is respectively 285.5, 302.5, and 317.9. This implies that two such machines using same processor each consume much more energy than a single machine executing the same work on two processors and shutting down the second one.

2.3 Web Service Modelling

Each job has its own behaviour and resource requirements. Often these requirements are not known in advance, so the system manages the amount of resources at each scheduling round taking as valid values the previous monitored demand. In other situations the user providing the job provides a guess on average or maximum requirements, and either the system trusts the (often inaccurate) user advice, or else over estimates resources. In this work we focus on web services, as their requirements are more sensitive towards amount of load than other high-performance tasks, and tend to change unpredictably in time. The methodology pro-posed in this work includes learning to predict, for each kind of job entering the system, the amount of required resources as a function of the kind of load it is receiving. When a new kind of job arrives we train a model mapping load to resources for that particular kind. A load is specified as a vector of load attributes determined beforehand (such as requests per second, bytes transferred per request or processing time per request). Given such a load, the model provides a vector describing the predicted usage of a number of resources (CPU, memory and bandwidth). Kinds of jobs could be web services running on specific software packages like Apache v.X, Tomcat v.Y, with attached modules like PHP or MySQL DB services. Each execution of the job provides pairs of workload attributes, resources used that can be used for (further) training the model de-scribing this kind of job.

2.4 Prediction on Scheduling

Once in the scheduling phase, the goal is to assign as few resources as possible to running jobs keeping user satisfaction but reducing power usage. This factor can be measured using the concept of Response Time (RT). This response time factor can be obtained prior to the arrival of workload, by monitoring client's requests and times for responses, but often the scheduler is interested in predict this information a priori.

2.5 Mathematical Model of the Datacentre

A grid based datacentre can be modelled as a set of resources, each one with a consumption cost, and a set of jobs to be executed with a set of resource requirements, profits and execution penalties. At each scheduling round what resources are assigned to each job must be decided, depending always in the requirements and conditions established by each SLA. The best solution will be that one that maximizes or minimizes a target function, usually describing the profit of the solution.

A proposed mathematical model for scheduling a binary matrix H x J, where H and J are the sets of (indexes of) hosts and jobs, and where each position [h, j] indicates whether job j is or not in host h. A valid solution must satisfy the condition that a job must be run entirely in one and only one host, as (by assumption at this stage) it cannot be split in different hosts. Each job needs a certain amount of resources to run properly at each moment, such as CPU quota, memory space and I/O access, predicted by the learned functions. Also each host has available a set of resources that cannot be overloaded, with an associated function relating usage to power consumption.

2.6 Schedule Solving

The functions of revenue, power cost, power, SLA are linear functions. The mathematic model to be solved by an Integer Linear Program (ILP) solver is well-known to be a NP complete problem, and therefore finite numbers of search algorithms are the only ones guaranteed to find the optimal solution. But heuristic approaches with often good performance are known, and also local-search methods can be very appropriate in a case as ours when we may have a fairly good solution from the previous scheduling round.

3 EXPECTED ENVIRONMENT

Expected configuration for the above mentioned methodologies is Intel Xeon 4 Core running at 3Ghz and with 16Gb RAM, running jobs of kind [Apache v2 + PHP + MySQL v5] in a virtualized environment [Ubuntu Linux 10.10 Server Edition + Virtual Box v3.1.8]. The simulated datacentre is expected to contain a set of 20 machines, representing copies of the model machine, each containing 4 processors 20 4CPUs.

4 CONCLUSION

Nowadays optimizing the management of datacentres to make them efficient, not only in economic values but also in power consumption, requires the automation of several systems such as job scheduling and resource management. And automation needs knowledge about the system to act in an intelligent way. As shown here machine learning can provide this knowledge and adaptively.

Also the effects of memory behaviour of web server platforms should be studied to model and predict the factors affecting memory occupation. Complementing the learned models revealing new hidden factors will make the model easier to handle and solve.

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 3, March 2015

REFERENCES

[1] Alexandra Carpen-Amarie, Anne-Cecile Orgerie and Christine Morin, Experimental Study on the Energy Consumption in IaaS Cloud Environments, 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing.

[2] Deepthi Dharwar, Srivatsa S. Bhat, Vaidyanathan Srini-vasan, Dipankar Sarma, Pradipta Kumar Banerjee, Approaches towards energy-efficiency in the cloud for emerging markets 2012 IEEE

[3] Shin-Jer Yang, Lee-Chung Chen, Hsi-Hui Tseng, Hui-Kuang Chung, Hsu-Yang Lin, Designing Automatic Power Saving on Virtualization Environment 2010 IEEE

[4] Josep Ll. Berral, Ricard Gavald'a, Jordi Torres, Adaptive Scheduling on Power-Aware Managed Datacentres using Machine Learning.