ISSN 2395-1621

Performance Evaluation and Dynamic Optimization of Speed Scaling on Web Servers in Cloud Computing

[#1]Saloni Bharne

[1]saloni.bharne@gmail.com

[#1]Computer Department

Savitribai Phule Pune University, Pune, India.

ABSTRACT

Cloud computing in the past few years has gained a lot of attention. This Internet-based computing provides shared computer processing resources and data to computers and other devices on demand. The main concern which arises is that due to its increasing demand, there have been significant concerns for cloud service providers such as increased financial costs of datacenters, large electricity demands, growing environmental pollution, etc. Thus, it is necessary to look out for the development of new energy efficient techniques suitable for datacenters. This can be achieved with the help of Stochastic Service Decision Nets (which is an extension of Stochastic Petri Nets) which investigates the energy efficient speed scaling on web servers. This is achieved by comprising Stochastic Petri Nets (SPN) along with Markov Decision Process model (MDP). This is done by maintaining two power states: low power state - when the load is light, server can be in low-power state to eliminate energy waste and high power state - high performance computing required when the load is high. The combined technique dynamically optimizes the speed scaling process and then makes performance evaluations.

ARTICLE INFO

Article History

Received: 25th March 2017

Received in revised form :

25th March 2017

Accepted: 28th March 2017

Published online :

31th March 2017

Keywords: Stochastic Service Decision Nets (SSDN), Markov Decision Process (MDP)

I. INTRODUCTION

Cloud Computing is a type of Internet based computing which uses a network of remote servers to manipulate, configure and access the application online. It provides shared computer processing resources, online data storage, infrastructure and application. Cloud Computing has gained huge popularity because of its platform independency i.e. it does not require any hardware or software installations on the local PC. Datacenter is a very important infrastructure component in cloud computing. It is like a building of networked computers along with data communications connections, backup power supplies, various security devices; basically all the computing infrastructure.

Demand for computing power and cloud computing technology is increasing day by day thereby datacenters end up consuming a lot of power and electric energy which not only affects the operational and maintenance cost but also leads to power dissipation issues leading to environmental problems. The major concern is that only 20-30% of the total server's capacity is utilized in a datacenter and the idle servers consume about 60% of the peak energy even when they are not doing any task. In cloud computing, both, the performance and the price of performance must be considered. Therefore, there is an urgent need to minimize this energy wastage and this can be done with the help of speed scaling on web servers.

In this paper, the speed scaling strategy on the web servers has been reviewed. Rest of the paper is organized as follows: Section 2 presents the speed scaling technique. Section 3 discusses the Stochastic Service Decision Nets model along with its working. Section 4 includes optimization techniques for the model described. Section 5 includes the conclusion.

II. SPEED SCALING ON WEB SERVERS

Speed scaling is a power management technique that involves dynamically changing the speed of the processor. Speed scaling is classified into static speed scaling and dynamic speed scaling. In static speed scaling the server runs only at a specified static speed whereas in dynamic speed scaling the server adapts the server speed based on the current server state.

Speed scaling on web servers will adapt the server speed to the workload. This means that when the load is light, the servers can be in a low power state with a little performance degradation and gradually when the load increases, the server can shift to a high power state where high performance computations are done. By making use of speed scaling on web servers, the unnecessary energy wastage can eliminated and servers can be utilized efficiently when high performance computing is needed. Speed scaling can be achieved by combining the SSDN and the MDP models.

III. STOCHASTIC SERVICE DECISION NETS PRELIMINARIES

A stochastic service decision net is defined as an 8-tuple Σ = (P, T, A, π, λ, Rt, Rm, F), where:

P = set of finite places

T = set of finite transitions

A = set of arcs

π = decision policy determining the transition for the decision maker

λ = set of firing rate for transition set T

Rt : T->R is a reward function when a transition T fires and Rm : M->R is a reward function when a marking is reached

F : objective function for decision maker

It is a service oriented modelling tool which means that it can capture major characteristics in various service patterns especially in cloud service scenarios. SSDN is an extension of Stochastic Petri Nets (SPN) which is a 5 tuple. It is a formal, graphical and executable method to describe complicated system behaviours. The SSDN model contains two basic structures: a decision maker sub-net and a system behaviour sub-net.

Consider a data center service model. When a user submits jobs, they first arrive at a front end proxy server and are distributed among the web servers with the help of load balancing. The jobs can be of type static content and dynamic content. The static content can only be fetched by the web server, whereas, the dynamic content requests accesses to the back-end data storage server which is in the form of databases in the hard disk. Constant requests to the hard disk slows down the system performance and the response time. For this, caches are maintained, which acts as a fast temporary storage, storing the most popular or frequently accessed data in them.

And while all of these activities are happening, the web servers are made to dynamically adapt the processing speed according to the workload. The figure given below, describes the working of the system with the speed scaling strategy. [1]

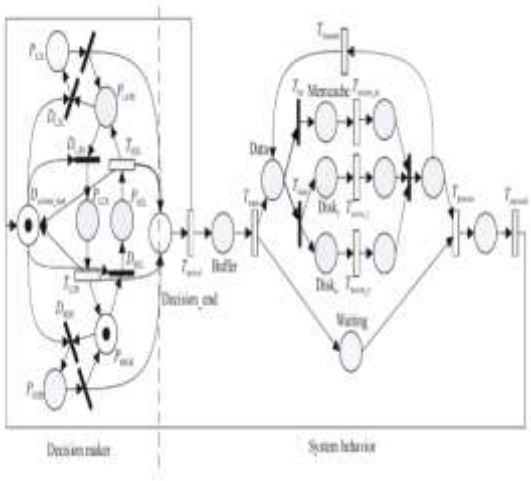

Fig 1. SSDN model

The decision maker sub-net makes decisions according to the system states and the system must react to these decisions with system behaviour.

The token in place Decision_start indicates that the decision maker has completed its final decision while the token in place Decision_end indicates that the decision maker has completed all its operations in the current round. Places P_{HIGH} and P_{LOW} denote that the server is either in high service rate providing good performance but with high power cost or it could be in low power state with some performance degradation with low power cost.

When the state is in high power state, i.e., when $\#P_{HIGH}$ = 1, the action space is $\{D_{H2H}, D_{H2L}\}$, where D_{H2H} indicates to maintain the high service rate in the current state and D_{H2L} indicates to transition into the low service rate. Similarly, for the low power state, i.e., when $\#P_{LOW}$ = 1, the action space is $\{D_{L2L}, D_{L2H}\}$, where D_{L2L} indicates to maintain the low service rate and D_{L2H} indicates a transition into high service rate. When the transitions happen to the alternative state, extra cost is incurred denoted by the time transitions T_{H2L} and T_{L2H}.

The jobs in the front-end proxy server will be routed to the waiting queue denoted by the Buffer. Once the job is ready to be executed, the Buffer will fire in parallel into the data access unit denoted by Data and into the execution waiting unit denoted by Waiting.

When the requested data is found to be present in the cache or hits the cache, it will access the cache memory denoted by the transition T_{hit}. If there is a T_{miss}, then more than one back-end data storage server will be accessed in parallel dented by T_{access_1} to T_{access_n}. When the data is ready the server will process with the detected service rate denoted by $T_{process}$. Finally, the job will successfully return its result to the end users.

IV. OPTIMIZATION TECHNIQUES

In order to reduce the complexity of the SSDN model and to cut down the number of states, four basic routing patterns have been included. They are parallel routing, selective routing, sequential routing, and iterative routing.

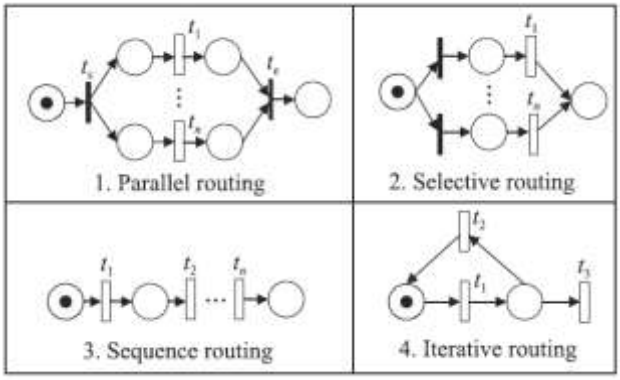

Fig 2. Routing patterns

- Parallel routing: In parallel routing, more than one job can be processed in parallel. This means that one token is forked into n tokens and these tokens fire transitions t_1-t_n in parallel.
- Selective routing: In selective routing, more than one transition can fire, but only one transition can fire at one time.
- Sequential routing: In sequential routing, the jobs are handled one at a time.
- Iterative routing: In iterative routing, a particular action executes continuously.

The Markov Decision Process model is a Continuous-Time Markov Chain (CTMC). It provides a mathematical framework for modelling decision making in situations where outcomes are partly random and partly under the control of a decision maker. The transition probability *Pr(s(t+1))* is defined as the transition from the state *s(t)* in time *t* to state *s(t+1)* in time *(t+1)* for action *a.*

The optimal policy is obtained by solving the MDP and then putting it into the SSDN, we get the performance and the energy metrics. These metrics are the response time and the saved energy. The response time is the length of the time taken by the system to react to an event and the saved energy is the probability of being in low service rate plus the energy gap between the high service rate and the low service rate.

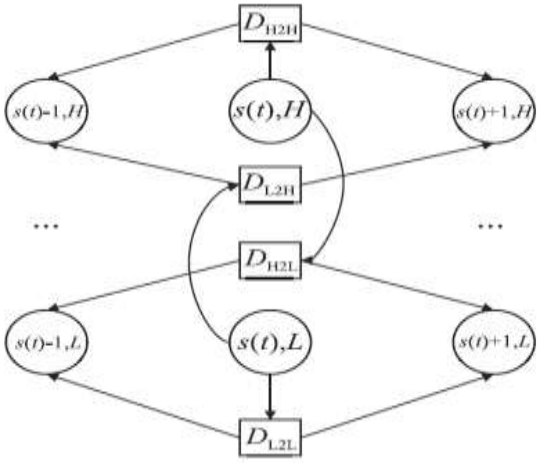

Fig 2. MDP model. Each state in this model is denoted by a 2-tuple as *{s(t),L/H}*, *s(t)* is the buffer size of place BUFFER and the *L/H* denoted whether in the low or high service rate state

V. CONCLUSION

With the increased use of cloud computing in the past few years, there has been a significant concern for the datacenters and its web servers consuming a lot of energy. Ideally, the web servers should consume the energy only during the peak state but the energy is also consumed even during the idle state leading to energy wastage. The Stochastic Service Decision Net (SSDN) and Markov Decision Process (MDP) model together investigate the speed scaling on web servers. With the help of MDP model, the decision maker sub-net of the SSDN model has been modelled to take decisions at the runtime in order to dynamically optimize the speed scaling strategy and provide performance evaluations and energy metrics. Thus, with the help of this model, the energy consumption can be managed by the datacenters efficiently by energy-efficient speed scaling on web servers and contribution to "green computing" is done.

REFERENCES

[1] Chuang Lin, Yuan Tian, Jianxiong Wan, Xuehai Peng and Zhen Chen, 'Performance Evaluation and Dynamic Optimization of Speed Scaling on Web Servers in Cloud Computing', ISSNll1007-0214ll09/11llpp298-307 Volume 18, Number 3, June 2013

[2] Ammar Rayes, Bechir Hamdaoui, Mehiar Dabbagh, and Mohsen Guizani, 'Toward Energy-Efficient Cloud Computing: Prediction, Consolidation and Overcommitment', IEEE Network (Volume: 29, Issue: 2, March-April 2015)

[3] Massoud Pedram, 'Energy-Efficient Datacenters', IEEE Transactions on computer-aided design of integrated circuits and systems, Vol. 31, No. 10, October 2012

[4] Jayant Baliga, Robert W. A. Ayre, Kerry Hinton, and Rodney S. Tucker, 'Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport

