An energy-aware approach for service performance evaluation

Alexandre Mello Ferreira
Dipartimento di Elettronica e Informazione
Via Ponzio, 34/5 - Milan - Italy
Politecnico di Milano
ferreira@elet.polimi.it

1. INTRODUCTION

Since the beginning of mobile device age, energy consumption has been affecting computers design. Nowadays it becomes also apparent in several areas previously led by computational performance. Recent IDC studies [5] have shown that around 16TWh of electricity was consumed by Western Europe servers in 2007 and it cost about 1.6 billion of Euros. Moreover it is expected that energy costs will exceed the hardware costs by 2015.

Thinking on this, sustainable computer resources development and usage could be applied to Information Systems (IS) by rearranging all IS resources (assets and capabilities) in order to accumulate, secure foster competitive advantages [7]. The main idea presented here takes into account energy consumption of servers to drive workload distribution, in such way that services can be dynamically routed to different servers. For that we consider the intrinsic data centre heterogeneity of servers, i.e. different computational power and energy consumption characteristics, classifying them from the slowest to the fastest at different energy consumption levels. Data centres servers heterogeneity combined with workload distribution aiming energy efficiency improvement is also described by [2], [1], and [4].

As each service has specific quality constraints to be satisfied, well described by Service Level Agreement (SLA), it is reasonable to think that some services do not need to be executed always into the fastest class of server. For example, a service responsible to generate the annual balance sheet does not need the same response time with respect to a service that concise stock in a trade report for the market department. Hence, services with non-critical performance constraints, such as response time, could be executed within slower servers to consume less power and to reduce their overall cost/price. Aiming to verify this dependency and to evaluate the impact of energy efficiency constraints in service centres, we present a system that is modelled as a closed queueing network with multiple-servers stations.

2. THE SYSTEM MODEL

In order to perform analytic evaluations, we use queueing network modelling approach which offers a reasonable balance between accuracy and efficiency. In this model, system resources (such as hardware resources and their software queues) are represented as network of queues and it is feasible to identify system bottlenecks. We also compute different performance metrics in simulated environments using Mean-Value Analysis (MVA) algorithm [6].

We start identifying the main components of the system (a typical configuration of a multi-tiered Internet service) and the interaction between hardware resources and software queues, which is done through Virtual Machines (VM). Each VM is represented as one server in our experiments and we assume the statement "One application - one server". This means that one VM can only be assigned to one service at a given instance of time. We present the first set of simulated results using Java Modeling Tools (JMT)[1], an open-source suite that supports common activities for performance evaluation.

By introducing the queueing network model above, we consider an open network system composed by N heterogeneous physical servers divided into three classes such that $N = \sum_{i=1}^{3} M_i$, in which M_i is the number of server at class station i. Moreover, the server utilization at class station i is given by U_i.

3. SIMULATION

The purpose of the simulation is to compare the differences between two different optimal network loads: (experiment A) considering only performance, which means, selecting always the server class with highest performance and highest energy consumption; (experiment B) taking into account of energy consumption of each class of servers, which means, selecting different classes through dynamic routing strategies. For both experiments we use an exponential distribution arrival rate (λ_{ir}) to model a Poisson process where the requests are independent from each other.

Agreeing with the model proposed, we assume that data centre controls the physical resource allocation issue using VM containers. Thus, the total physical server energy consumption is split properly among its VMs based on components utilization rates. The evaluated scenario is a typical multi-tier web system model composed by a set of multi-class

[1] `http://jmt.sourceforge.net`

multi-server queues grouped as web servers (WS), application servers (AS), and database servers (DS) and divided into classes according to their characteristics. Considering that many data centres are over-provisioned for peak usage, the average scenario allows us to play with such new energy efficient configuration without reducing service quality [1].

Regarding to the queueing policies, all servers support infinite requests in queue (actually, it is possible due to admission control mechanism which takes care of this matter) and requests are executed according to First Come First Served (FCFS) queueing discipline. In such rule the executions comply with the request arrival order.

Experiment A - based on performance optimization The first experiment is conducted aiming only performance optimization with respect to throughput and response time. The routing strategy sets class-r services for the smallest queue length stations. Preliminary results have shown that none of the servers has utilization rates bigger than 38%, which means that there is some space to redistribute the workload considering energy consumption values without bottlenecks. The goal is to change the routing strategies among servers in such a way to increase the utilization of low energy consumption servers (i.e., till burst lower bound) while keeping the high performance servers into energy saving mode.

Experiment B - based on energy awareness. The second experiment focuses on energy consumption and, for that end, the routing strategy is changed for a dynamic probabilistic one. Unfortunately JMT does not provide any kind of routing strategy based on either server energy consumption or hardware characteristics. Due to that limitation, we defined some reasonable probabilities values to represent possible energy consumption instant pictures of the system. In the best energy consumption scenario, the workload is routed mainly based on the server energy consumption rate. Considering the dynamic routing of real situations, other scenarios show some new situations in which the first one is not possible due to slow servers capacity. Therefore the system dynamically changes the probability of routing in order to avoid slow servers burst situation. The burst situation is represented as a queue length limit based on the server class.

4. ENERGY EVALUATION

In order to compare both experiments from energy point of view, we do need to put them in terms of energy consumption. For that, all servers utilization values are mapped into the server class and their utilization modes together with the number of servers in a specific class of server. we sum up the amount of energy consumed by each class of server according to their utilization mode and multiply it by the number of servers into that situation. The formula for this calculation is expressed in (**1**). The function f_i finds out the respective amount of energy consumed in each execution period.

$$ec = \sum_{i=1}^{I} f_i\left(U_i\right) \cdot M_i \qquad (1)$$

The idea presented here is the part of a bigger research work, that combines different energy-aware solutions toward energy efficiency at service level. In fact, this approach fits with the work introduced by [3] which obtains a better trade off between performance and energy efficiency of composite services considering non-functional characteristics.

Nevertheless, we consider only energy consumption of servers in a simulated environment and further investigations have to be done to evaluate their impact over cooling (which represents about 50% of the total facility energy consumption) and power infrastructure. Reducing servers energy consumption has a immediate impact into the whole data center operational costs (specially the electricity bill), but also indirectly throughout economic incentives programs launched by stakeholders or government.

5. ACKNOWLEDGMENT

This work was partially supported by funding from the Information Society Technologies Programme under the 7th Research Framework Programme of the European Union, within the research project GAMES (Green Active Management of Energy in IT Service centres). Call: FP7-ICT-2009-4 Grant agreement no.: 248514

6. REFERENCES

[1] J. Burge, P. Ranganathan, and J. L. Wiener. Cost-aware scheduling for heterogeneous enterprise machines (cash'em). In *CLUSTER '07: Proceedings of the 2007 IEEE International Conference on Cluster Computing*, pages 481–487, Washington, DC, USA, 2007. IEEE Computer Society.

[2] X. Fan, W.-D. Weber, and L. A. Barroso. Power provisioning for a warehouse-sized computer. In *ISCA '07: Proceedings of the 34th annual international symposium on Computer architecture*, pages 13–23, New York, NY, USA, 2007. ACM.

[3] A. M. Ferreira, K. Kritikos, and B. Pernici. Energy-Aware Design of Service-Based Applications. In *ICSOC & ServiceWave 2009: Proceedings of the 7th International Joint Conference on Service-Oriented Computing*, Stockholm, Sweden, 2009. Springer.

[4] R. Nathuji, C. Isci, and E. Gorbatov. Exploiting platform heterogeneity for power efficient data centers. In *ICAC '07: Proceedings of the Fourth International Conference on Autonomic Computing*, page 5, Washington, DC, USA, 2007. IEEE Computer Society.

[5] G. Nebuloni and T. Meyer. The role of x86 processors in the virtualized, energy-constrained datacenter. Technical report, An IDC Multiclient Study, November 2008.

[6] M. Reiser and S. S. Lavenberg. Mean-value analysis of closed multichain queuing networks. *J. ACM*, 27(2):313–322, 1980.

[7] N.-H. Schmidt, K. Erek, L. M. Kolbe, and R. Zarnekow. Towards a Procedural Model for Sustainable Information Systems Management. In *HICSS' 09: Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10, Hawaii, USA, 2009. IEEE Computer Society.