Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 318 (2015) 159–177

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Operating Policies for Energy Efficient Dynamic Server Allocation

Thai Ha Nguyen, Matthew Forshaw and Nigel Thomas

School of Computing Science, Newcastle University, UK

Abstract

Power inefficiency has become a major concern for large scale computing providers. In this paper, we consider the possibility of turning servers on and off to keep a balance between capacity and energy saving. While turning off servers could save power, it could also delay the response time of requests and therefore reduced the performance. Furthermore, as consistency is one of the most important factors for a system, we also analyse the level of consistency in the form of switching rate and fault occurrence. Several heuristic-based switching policies are introduced with a view to balance the cost between power saving, performance and consistency. Simulation results are presented and discussed with requests arriving according to a two-phase Poisson process.

Keywords: Energy efficiency, discrete event simulation, performance evaluation.

1 Introduction

The non-functional challenges facing large scale computing provision are generally well documented [13]. Amongst these the cost of energy has become of paramount concern. Energy costs now dominate IT infrastructure total cost of ownership (TCO), with data centre operators predicted to spend more on energy than hardware infrastructure in the next five years. The U.S. Environmental Protection Agency (EPA) attribute 1.5% of US electricity consumption to data centre computing [4], and Gartner estimate the ICT industry was responsible for 2% of global CO_2 emissions in 2007 [17]. With western european data centre power consumption estimated at 56 TWh/year in 2007 and projected to double by 2020 [3], the need to improve energy efficiency of IT operations is imperative.

Data centres, with their high density of power consumption and a steady growth in number, have become a major industrial energy consumer in the recent years. One of the most important factors that promoted their growth is that cloud computing

[1] Corresponding author: nigel.thomas@ncl.ac.uk

http://dx.doi.org/10.1016/j.entcs.2015.10.025

has become a big trend in web services and information processing. The most significant advantage of the cloud is its flexibility. It offers the chance of shifting capital expenditure to operational expenditure [7], which is ideal for starting a new service. Furthermore, since there is an increase in the quantities of data being collected for commercial, scientific or medical purpose, the big capacity of data centre is ideal to process such massive volume of data. As the cloud offers users an illusion of infinite computing resources on-demand [9], cloud computing is in fact essential to gather useful data from that enormous amount of information [1].

One of the more challenging problems in managing energy consumption in distributed systems is in handling variability of workload. There are a number of measures which can be applied to manage the effect of variable supply and demand. For example, there are a variety of load balancing techniques [11] and traffic shaping measures [12] which can be utilised to manage demand so that resources do not become excessively over-utilised when demand is high. An alternative approach is to dynamically manage the supply of service capability by making more servers availible during periods of high demand. Slegers *et al* [18,19] considered the problem of finding the optimal share of servers to different services under variable load in order to minimise a performance-based cost function.

This paper is based on the work of Slegers *et al* [20] and is focussed on the notion that servers can be powered off and on according to demand in order to avoid the non-trivial energy requirements of idle servers. With perfect knowledge of arriving workload an optimal dynamic allocation of servers can be obtained which significantly reduces the overall energy demand of the system with no impact on performance, i.e. servers can be made available only when they are going to be used. Of course, we do not generally have a perfect knowledge of future workload and so an optimal dynamic solution is not practical. Instead we must investigate the trade-off between energy consumption and performance (e.g. response time) to determine the best practical method of reducing energy costs whilst not adversely affecting the quality of service. Two principle approaches to minimising energy consumption are apparent. In the first instance an optimal fixed provision of servers can be computed based on estimated workload. Depending on the variability in demand, this approach might lead to servers being idle for extended periods or to some tasks experiencing long waiting times during peak demand. The second approach is to compute a strategy to turn servers on and off based on the current (or past) state of the system. This approach minimises idle time by turning off servers, but potentially delays tasks which arrive in a burst as it takes time to turn servers back on. In addition, powering servers off and on may lead to faults which not only reduce the total available number of servers, but may also further delay an arriving task.

The remainder of this paper is organised as follows. In the next section we explain the context of this work in relation to other work on energy reduction. In Section 3 we describe the system model and introduce six heuristic strategies for controlling the number of servers powered on and off. This is followed in Section 4 by a brief description of the simulation environmen and we then a present and discuss the results of our experiments. Finally we present some conclusions and