Ad Hoc Networks 25 (2015) 554–564

Contents lists available at ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc

Analysis of an energy proportional data center

Ricardo Lent

University of Houston, TX, United States

ARTICLE INFO

Article history:
Received 26 March 2014
Received in revised form 12 September 2014
Accepted 3 November 2014
Available online 11 November 2014

Keywords:
Energy proportionality
Computing cluster
Energy efficiency
Power
Quality of service
Service level agreement

ABSTRACT

Energy proportionality is a desirable property of an energy efficient data center that can be achieved by making servers available on demand, dynamically enabling enough computing capacity to handle the workload. However, reducing the number of running servers can impact job performance and may potentially lead to breaches in the service level agreement. We analyze the optimal (minimum) energy requirement of servers in an energy proportional data center to maintain a selected performance service level objective from the following possibilities: (i) running servers at or below a maximum utilization level; (ii) keeping the average job response time below a given limit; and (iii) limiting the probability of job response times exceeding a turnaround deadline. Performance and power measurements from a real server allow to define realistic parameters for theoretical and simulated models and to obtain realistic results.

1. Introduction

Data centers consume large amounts of electricity-currently about 20–30 gigawatt worldwide [1], of which a large proportion is wasted by idle computers. The rationale for leaving computers idle is often justified by service level agreements that require operators to deliver services within certain limits. These idle computers then remain waiting for a possible increase in user demand to handle the computing requirements of the additional workload while maintaining the quality expected by the users. As a result, a typical data center has a peak utilization of only 40% with long low-demand periods, some of which with utilization levels as low of 5% [2]. In addition to the power needed to handle workload, the idle power also contributes to increase data center heat and in rising cooling costs [3].

Energy proportionality is a simple concept that can help to boost energy efficiency. The idea is to dynamically manage capacity, so that excess resources and their energy consumption, can be temporarily removed from the system, and restored later when needed. The concept has been already implemented, in various forms in computing systems [4]. Many central processing units can adjust their operating frequency and voltage to reduce electrical energy usage during low-activity periods. Storage devices, such as, hard drives, can stop spinning when not used for some time. Server and data center cooling also consumes an energy proportional to the heat that is removed. In spite of these energy saving mechanisms that make computers proportional, server hardware is still far from the ideal vision of the energy proportional computer and still draw considerable power when idle [5,6]. In most cases, the usual resource redundancy available by design in data centers, should make possible the implementation of energy proportionality by dynamically managing the power state of the machines without affecting service levels.

In this paper, we investigate the optimal energy consumption of the computing cluster in a proportional data center that is required to maintain a specific service level objective. The study assumes an abstract reasoning with servers that run job requests without referring to any particular computing application. Implementation details involving the assignment or movement of workload from

http://dx.doi.org/10.1016/j.adhoc.2014.11.001
1570-8705/

a group of servers to another would depend on the specific computing model used in the cluster. We will keep the discussion generic, but it should be clear, that the analysis is applicable to many computing models, including high performance computing and cloud computing, for example, through virtual machine migration in a PaaS environment. The analysis allows to gain insight into the energy and performance tradeoffs when specific performance objectives must be satisfied while keeping the server provisioning proportional to the user demand.

2. Data center model

The large number of servers commonly found in data centers actively contributes to their total energy consumption. In addition, cooling requires an energy that among other parameters, is proportional to the one dissipated by these servers, incrementing the total energy consumption figure by a factor provided by the *power usage effectiveness* (PUE) metric–a characteristic value of data center deployments. The power consumed by servers depends on their operating state and the level of workload being handled. Let us assume that a data center consists of n computing nodes (servers) and that a given time, only m of them are running ($m \leqslant n$) and the rest $n - m$ are in hibernation.

2.1. Power model

Assuming homogeneous servers and ideal load balancing among the nodes that are running, the power consumption of the computing cluster that is handling jobs that arrive at a rate Λ can be calculated by:

$$\Pi(\Lambda) = n\kappa(I + J\rho) + n(1-\kappa)H + O;\ \kappa <= 1 \tag{1}$$

where $\kappa = m/n$ is the ratio of running to hibernating nodes, I is the idle power consumed by each server, j is the power increment due to utilization ρ, H is the node power consumed while in hibernation, and O is the power consumed of other equipment in the facility, such as network switches [7], uninterruptible power support, and network storage equipment.

We have made a number of assumptions to keep this model simple. Servers are assumed to be of identical characteristics, with identical idle and operating power under the same load. We have assumed a single common state for non-running servers. That is, servers may hibernate, but not sleep or be switched off. Another assumption is that the power consumed by other equipment is constant. It should be clear that the first two assumptions can be easily relaxed by splitting the first two terms of the right-hand part of Eq. (1) to consider cases for each server type in the network. The power consumed by other equipment may not be constant in a real data center. However, we could think of their power consumption as consisting of two elements: a constant part being represented by parameter O in Eq. (1) dynamic part that can be added to parameter J.

Because of the assumption of homogeneous servers, the optimal job allocation occurs by uniformly distributing the job arrival rate Λ among the running servers. Also, we assume that we know the characteristics of the workload, so that jobs take on average $E[S]$ seconds to complete on a single machine. Including the waiting time for service, jobs take on average W seconds to complete.

2.2. Service level objectives

A service level agreement (SLA) involve the acceptance by both the provider and its customers of different aspects of a service delivery. These service aspects can include performance levels, customer obligations, problem and disaster recovery, etc. We are mostly interested in the quantitative performance aspects of SLAs, which require data centers to offer services at, or at a better level, than a given measurable service level objective (SLO). In relation to evaluating a data center operation, we consider three kinds of SLOs that can be directly studied with proposed model.

Most performance-related service level objectives can be fulfilled by limiting the utilization level of servers. The first SLO (SLO-0) does just that, requiring jobs to run in servers that have a utilization level of at most ρ_M:

$$\rho_i \leqslant \rho_M;\ i = 1, 2 \ldots, n \tag{2}$$

where ρ_i is the utilization of the ith server and n is the number of servers in the data center as before.

Response time, which is the time interval between the initiation of a job request and the completion of the job, is another important performance metric regarding the responsiveness of a service and plays a significant role in user experience. It is commonly used in industry as a SLO metric [8]. A second SLO (SLO-1) constraints job requests to complete on average within α times the average service time:

$$W \leqslant \alpha E[S];\ \alpha \geqslant 1 \tag{3}$$

where W is the average job response time for the system and $E[S]$ is the average job service time. Parameter α defines the level of quality tolerance allowed for the service. It is sensible to make the response time limit a function of the average service time of the workload given that clearly, the former cannot be lower than the latter (i.e, α cannot be less than 1). The value of $\alpha E[S]$ gives the maximum permissible average response time for jobs, and longer response times than $\alpha E[S]$ are a violation to the SLO-1. Intuitively, we should expect an increase in the rate of SLO violations, if we aggressively reduce the number of running servers in the data center regardless of the workload level. Having less computing resources available to handle demand, will likely result in jobs waiting longer for service. At the same time, lower power consumption should be expected after forcing some servers to hibernate, i.e., by decreasing κ in Eq. (1).

While SLO-1 constraints the average job response time, another important service level objective commonly used in industry and that is not well captured by the SLO-1 definition is the *maximum response time* for any job [8,9]. The third SLO (SLO-2) models this important metric by setting a limit to the probability of jobs exceeding the maximum response time, that is, by requiring each job to complete within a time bound $B : w_i \leqslant B$, where w_i is the response