Sustainable Computing: Informatics and Systems 3 (2013) 80–93

Contents lists available at SciVerse ScienceDirect

Sustainable Computing: Informatics and Systems

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

A model for network server performance and power consumption

Ricardo Lent

Imperial College London, United Kingdom

ARTICLE INFO

Article history:
Received 29 May 2011
Received in revised form
14 November 2011
Accepted 2 March 2012

Keywords:
Green computing
Power consumption
Data center
Multiclass queuing network
Performance evaluation

ABSTRACT

We investigate the power consumption of a network server based on a multiclass open queuing network with closed-form solution. The model consists of service stations that represent major subsystems within the server. We also propose a practical methodology for fixing model parameters through a set of simple tests. The resulting model can reasonably predict the power usage trends with respect to workload intensity that can be observed in real computing systems. Supporting experiments provide a cross validation of the model against performance and power measurements obtained on three types of commercial systems.

1. Introduction

The increasing need for computing power in data centers brings along concerns about their high electrical energy consumption. Something similar occurs in the Internet where the increasing interest of users for digital media and various kinds of online social services that can be consumed from a plethora of end-user devices creates a driving force that continuously pushes service providers to expand their computing capacity and increase their energy consumption. A capacity increase may involve adding new servers or replacing servers with machines of better specifications, and may as well involve deploying new data centers to house thousands of new servers.

Each data center can consume many megawatts of electrical power, making data centers responsible for the use of about 1.5–2% of the global electricity. This figure is growing at a fast rate (12% per year) according to a recent report [1] and in agreement with trends previously forecasted [2]. Depending on the efficiency of a data center, servers will contribute to 50–83% of the total power consumption. These values correspond to a Power Usage Effectiveness (PUE) of 1.9 (normal value) and 1.2 (for a state-of-the-art data center). The remaining energy is consumed by other functions, such as cooling [3,4].

In this context, the emergence of green computing comes at the right time as more efficient practices are needed to deal with the increasing demand of electricity. The challenge is to reach at least the same level of service but with the lowest amount of energy [5]. A central piece that is needed for the development of sustainable computing systems is a proper understanding of the power consumption dynamics of servers.

We propose a model that can be employed to make predictions of a server's power consumption and performance. The model is simple enough to be computationally efficient, and therefore, suitable to a wide range of applications, but also accurate and flexible to accommodate "what-if" type of studies.

The contributions of this paper are:

- A multiclass queuing network-based model that approximately replicates the way major power consumers within a server system interact under heterogeneous workload. The activities of these power consumers can be easily monitored in practice, so model parameters can be quickly derived. These subsystems are characterized by the utilization of CPU cores, disk drives, and network ports.
- A power model for each subsystem that can collectively estimate the power consumption of the server under a sustained workload.
- A methodology to derive model parameters from a simple set of tests that can be easily applied to different types of servers.
- Finally, we introduce a calibration improvement to the Electrical Power Usage Monitoring System (EPUMS)—a sensor network developed by the author in a previous work [6]. We use EPUMS to illustrate the experimental acquisition of parameters for real server models and compare predictions to a large set of measured values.

2. Related works

The majority of power consumption models in the literature that are applicable to servers offer the advantage of simplicity but lack

E-mail address: r.lent@imperial.ac.uk

2210-5379/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.suscom.2012.03.004

accuracy as decided by Rivoire et al. [7] after examining 5 representative full-system power models in a recent study. On the other hand, most power models have been developed with a black-box approach in mind where power consumption is matched to a server's load (usually CPU utilization) from measured data. Usually a single parameter is used although more than one has been used in some works, such as in the case of Lewis et al. [8], who applied a linear regression model to observations of power consumption in server blades with respect to various metrics. In another work, Sasaki et al. [9] investigated linear models for power and performance in web server clusters depending on the load of CPUs and disks, but left open the explicit calculation of these loads.

These models tend to produce good predictions of power consumption only as long as the characteristics of workload do not change. This strong dependency on workload creates also the non-trivial problem of how to apply these models to different hosts or after changes are made to the original system. Moreover, the power consumption of future systems is expected to become less CPU-dominated due to the availability of faster and larger memory and storage systems, which could reduce the applicability of these models to future systems.

Another problem with existing models is that they tend to be linear, which limits their usefulness to partial utilization regions and server types. It has been observed that the power consumption of a server exhibits different patterns depending on its utilization level. A single linear model will therefore introduce large prediction errors to certain regions. This observation is consistent with the work of Chia-Hung et al. [10] who suggested that a linear model of power consumption could be accurate only up to medium utilization levels. Examples of linear power models abound: Chu et al. [11], Jaiantilal et al. [12], Yuan and Ahmad [13]. Some measurement studies have suggested the suitability of linear models to some extent. Fan et al. [14] obtained measurements of the power consumption of warehouse-sized computers (computer for large-scale Internet services), Li et al. [15] did a similar work on web servers running on blade systems, and Economou et al. [16].

Similar models have been applied to guide the design and in some cases operation of power optimization solutions. For example, such models were used by Brochard et al. [17], who proposed a power-aware processor scheduling mechanism, Nedevschi et al. suggested power reductions through idle states [18], Gandhi et al. [19] studied optimization techniques for server farms, whereas Elnozahy at al. [20] focused on clusters. The literature is rich in power reduction techniques for computing systems. A comprehensive survey of green networking research was compiled by Bianzino et al. [21]. Another survey on the power and energy management in servers was done by Bianchini and Rajamony [22]. Some relevant examples of power optimization techniques are the works of Sankar et al. [23] on metric composition energy-delay, Chase et al. [24], who proposed an economic approach to server resource management. Rodero et al. researched application-aware power management looking at individual components [25]. A popular approach to reduce power usage is by switching off unused equipment as done by Chen et al. [26] and Niyato et al. [27]. A control mechanism to adjust the peak power of a high density server was suggested by Lefurgy et al. [28] by means of a feedback controller.

The simplicity of this kind of models allows their use in high demanding applications. A good example is the work of Meisner et al. [29] who proposed the use of a linear PSU model with a low-order FIR filter to predict the peak power of data center servers based on real-time observations of CPU utilization. The model was kept simple to allow easy integration into a Linux kernel that may lack floating-point support. Another set of works have addressed the power of specific components. Vogelsang [30] proposed a model to calculate energy consumption of DRAM by technology, which should allow forecasting figures for generations of devices yet to come. Server's memory may contribute in a significant proportion to the total system's power consumption. Deng et al. proposed the creation of active low-power modes for DRAM ranks [31] as a way to reduce the energy consumption of memory. In another approach with a similar aim, David et al. introduced a control algorithm to adjust both memory voltage and frequency depending on memory bandwidth utilization [32]. Some works have also been devoted to study the power consumption of virtualized applications [33,34]. Meisner et al. proposed a sophisticated model of power consumption for the PowerNap server architecture [35] that used a M/G/1 queuing system to derive predictions.

Proper instrumentation is also needed to guide the development and definition of power consumption models. Few works in the literature have been devoted to develop the tools necessary to measure power consumption in computing systems. The work of Bedard et al. [36] aimed to develop a measurement system for the internal devices within a computer system, whereas Kansal and Zhao [37] discussed the development of fine-grained automated tools to profile energy usage of computer resources looking to embed such a system to application development and profiling. Lent [6] developed a sensor network for high-frequency power measurements suitable from few to large number of computing and networking devices.

3. Model

3.1. Overview

The model consists of a non-blocking and multiclass open queuing network [38–40] containing 3 different types of service centers. A service time characterizes each center: $\mathbf{S} = \{\mu_C^r(0)^{-1}, \ldots, \mu_C^r(C-1)^{-1}, \mu_D^r(0)^{-1}, \ldots, \mu_D^r(D-1)^{-1}, \mu_N^r(0)^{-1}, \ldots, \mu_N^r(N-1)^{-1}\}$, where $r \in \{0, 1, \ldots, R-1\}$ assuming R different job classes in the system. In the notation, superscripts indicate a job class and the index between parenthesis enumerates same-type centers. Subscripts indicate the type of service center. Job passages of class r between service centers are modeled by transition matrix $\mathbf{P}^r$. External job arrivals are defined by vector: $\mathbf{L}^r = \{\bar{0}, \Lambda^r(0), \ldots, \Lambda^r(N-1)\}$, where vector $\bar{0}$ is of length $C+D$.

Service centers may have associated an infinite queue and are assumed to be FCFS with per-class service rates but independent of the load. Each center represents a given subsystem within a computer that is expected to be active during service delivery. A subsystem is an abstract representation of hardware and software elements present in a server. However, specific components within each subsystem will be used to characterize their entire utilization, given that it would be infeasible to monitor all and each computer component to derive specific models of utilization. Nevertheless, the utilization of certain components is relatively easy to measure and a certain level of correlation in the activities of different elements is to be expected.

We consider three types of subsystems: core, drive, and network port, whose utilizations can be easily measured in practice. A given system will contain C cores, D drives, and N ports. The service rates defined above correspond to: $\mu_N^r(i)$ for network ports ($i \in \{0, \ldots, N-1\}$), $\mu_C^r(j)$ for cores ($j \in \{0, \ldots, C-1\}$), and $\mu_D^r(k)$ for drives ($k \in \{0, \ldots, D-1\}$) for class r.

A core subsystem models the whole activity of a single CPU core and some activities of cache memory, main memory, and supporting buses that are expected to be used while the core is active. For simplicity, we do not model memory as a separate subsystem, but assume memory to be integral part of other subsystems.

The drive subsystem can represent either a single hard disk or a disk array depending on the modeling needs. The network port