Ad Hoc Networks 25 (2015) 497–504

Contents lists available at ScienceDirect

Ad Hoc Networks

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

Automatic energy efficiency management of data center resources by load-dependent server activation and sleep modes

Paul J. Kuehn [a,*], Maggie Ezzat Mashaly [b]

[a] Institute of Communication Networks and Computer Engineering, University of Stuttgart, Pfaffenwaldring 47, 70569 Stuttgart, Germany
[b] Networks Department, German University in Cairo (GUC), Egypt

ARTICLE INFO

Article history:
Received 26 June 2014
Received in revised form 11 November 2014
Accepted 11 November 2014
Available online 20 November 2014

Keywords:
Data center server consolidation
Resource management
Energy efficiency
Service Level Agreement (SLA)
FSM-controlled queuing model
Performance evaluation

ABSTRACT

The operation of large Data Centers (DC) with thousands of servers is very costly in terms of energy consumption and cooling requirements. Currently, major efforts can be observed for server virtualization and consolidation to approach a proportionality between computation amount and energy consumption. In this contribution, a generalized model is presented which allows an automatic server consolidation by a load-dependent control of server activations using multi-parallel hysteresis thresholds, cold and hot server standby, and Dynamic Voltage and Frequency Scaling (DVFS). For the energy-efficiency and performance analysis, a multi-server queuing model is defined which is controlled by a Finite State Machine (FSM). The parameters of the queuing model are defined such that Service Level Agreements (SLA, e.g. as mean or percentiles of response times) are guaranteed except for overload conditions. The queuing model can be exactly analyzed under Markovian process assumptions from which all relevant quality of service (QoS) and energy efficiency (EE) metrics are derived. Numerical results are provided which demonstrate the applicability of the proposed model for the DC management, in particular to theoretically quantify the tradeoff between the conflicting aims of EE and QoS.

1. Introduction

High-speed fixed, mobile and wireless networks, web-based, social and multi-media applications are the driving forces of the currently ongoing paradigm shift from a communication-centric to an information-centric internet. Traditional IT infrastructures are challenged by cloud networks with distributed data centers which make full use of multi-core processing and huge storage capacities allowing fast access to data, search and business process applications. The huge energy demand of these data centers contributes significantly to the energy consumption and the "carbon foot print" and caused world-wide efforts to the "Green ICT" movement.

Energy efficiency can be improved on quite different levels: device level by a steady miniaturization of transistor elements, on the circuit/chip level by low-power circuit design, on the network level through efficient use of bandwidth by modulation and coding, switching and routing protocols, on the application level by energy-aware user behavior, and on the systems level by system operation management. This paper addresses energy efficiency management, specifically through modeling and quantitative analysis by stochastic queuing theory. In this approach, energy-consuming resources are modeled as "servers" (e.g., a processing device which executes a task) and "buffers". The energy consumption of a processor is usually simplified by a constant representing the average consumption during processing. Variable energy consumption

* Corresponding author.
E-mail address: paul.j.kuehn@ikr.uni-stuttgart.de (P.J. Kuehn).

http://dx.doi.org/10.1016/j.adhoc.2014.11.013
1570-8705/

through, e.g., dynamic pipeline operations or caching, belong to a lower level closer to the hardware and require more knowledge on the underlying application program and will not be considered in this contribution. The dynamic behavior of a whole data center or of specific server groups is modeled by "stochastic arrival processes" of processing tasks and task execution times are modeled by "stochastic service times". Tasks which cannot be executed immediately are buffered in a "queue".

Actions for energy efficient operations on the systems level are deactivations at low-load situations and activations at high-load situations, sleep mode (non-operative mode, e.g. by lowering the power supply to avoid booting in case of reactivation), or slow-mode operation by clock frequency throttling; the latter two operations are known as Dynamic Voltage and Frequency Scaling (DVFS). Control actions on multiple servers aim at resource management through adaptive assignment of tasks to server groups by virtualization methods, server consolidation by activity monitoring and deactivation of servers, by load sharing for tasks among different processors (scheduling), or by load balancing through shifting tasks to other virtual machines of the same or of remote data centers (task or process migration).

Energy efficiency has become a hot research topic in the recent years reflected by high publication activities. Most contributions address architecture, measurement and management issues, c.f. [1–4]. Another group of papers approach the problem by modeling and queuing theory, see, e.g. [5], where the data center is modeled by a multi-server queuing system. Self-adaptive server consolidations have been suggested and modeled by the authors on the basis of hysteresis mechanisms by FSM-controlled queuing systems [6–8] and applied to load balancing [13]. The current paper is a revised and extended version of a conference paper [14]; it presents a more detailed modeling approach for multi-server queuing systems which are controlled by a Finite State Machine allowing automatic adaption to the load level and a detailed consideration of specific control schemes and overhead involved with dynamic server activations.

The rest of the paper is structured as follows: In Section 2, the structure and the parameters of the queuing model are presented together with the design criteria for the intended operation mode of the FSM which controls the adaptive algorithms for power saving. In Section 3, a short review is given over existing literature on queuing models with hysteresis control. For the generalized model, we have developed a new recursive solution algorithm which allows for an effective calculation of the probabilities of state for an arbitrarily large number of servers and the most characteristic performance and energy consumption values. Finally, Section 4 provides a numerical case study and discusses the parametric influences on the performance.

2. Queuing model

According to the stated features of the DC queuing model, we have to construct the State Transition Diagram (STD) of the FSM in a systematic way. For this, the following characteristics have to be satisfied by the FSM:

(1) Multiple hysteresis thresholds to avoid frequent oscillations between activations and deactivations of server resources to serve stochastically varying service requests and for an automatic self-adaptation to highly volatile load variations.
(2) Throttling of new server activations upon short load bursts by buffering of the requests up to scalable upper thresholds.
(3) Serving of task requests with the maximum service rate of the activated servers to keep delays as small as possible as long as (4) is not affected.
(4) Threshold parameters of the hystereses have to be set such that a prescribed Service Level Agreement (SLA) is guaranteed except for overload situations. Overload situations are defined when all servers are already activated and new requests could not be served under the given SLA.
(5) Throttling of server deactivations when the queue of waiting task requests falls below of a scalable lower threshold (implemented by the DFS principle).
(6) Consideration of two different deactivation modes:
    (6.1) If a server becomes idle, it will be set in a sleep mode with lower power consumption from which it can be reactivated quickly ("warmup") without booting ("hot stand-by mode", HSB).
    (6.2) If a server becomes idle, it will be completely deactivated (switched-off) and has to be booted again if a new server activation is required ("cold stand-by mode", CSB).
(7) Sleeping or deactivated servers are activated again at the instant of a task request arrival and under a predefined threshold for buffered requests.
    (7.1) In case of a sleeping server, activation takes a short warmup time.
    (7.2) In case of a deactivated (switched-off) server, activation requires a longer activation time for booting. Activated or reactivated servers start servicing buffered requests immediately after booting or warmup according to the FIFO (First-In, First-Out) queue discipline.

STDs for multiple serial and parallel hystereses without activation overhead and without DVFS were reported in [6,7] and extended to activation overhead in [8] by the authors. This paper extends these results with respect to DVFS, cold and hot stand-by.

Fig. 1 shows a generic queuing model with dynamic activations/deactivations of servers acc. to [6–8] for generally distributed task requests (Arrival Process Type G, arrival rate λ tasks/s), generally distributed task service times (Service Process Type G, maximum service rate μ per activated server), buffer with capacity s, the Finite State Machine (FSM), a server group with n servers, and a scheduler. The state of the queuing system is indicated by the vector (X, Z), where X denotes the current number of busy (i.e., service executing) servers and Z of waiting task requests. The variables X and Z are reported to the FSM