Sustainable Computing: Informatics and Systems 4 (2014) 292–306

Contents lists available at ScienceDirect

Sustainable Computing: Informatics and Systems

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

Energy-efficient and thermal-aware resource management for heterogeneous datacenters

Hongyang Sun*, Patricia Stolf, Jean-Marc Pierson, Georges Da Costa

IRIT, University of Toulouse, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France

ARTICLE INFO

Article history:
Received 1 March 2014
Received in revised form 7 July 2014
Accepted 11 August 2014

Keywords:
Datacenter heterogeneity
Online scheduling
Server placement
Cooling
Multi-objective optimization

ABSTRACT

We propose in this paper to study the energy-, thermal- and performance-aware resource management in heterogeneous datacenters. Witnessing the continuous development of heterogeneity in datacenters, we are confronted with their different behaviors in terms of performance, power consumption and thermal dissipation: indeed, heterogeneity at server level lies both in the computing infrastructure (computing power, electrical power consumption) and in the heat removal systems (different enclosure, fans, thermal sinks). Also the physical locations of the servers become important with heterogeneity since some servers can (over)heat others. While many studies address independently these parameters (most of the time performance and power or energy), we show in this paper the necessity to tackle all these aspects for an optimal resource management of the computing resources. This leads to improved energy usage in a heterogeneous datacenter including the cooling of the computer rooms. We build our approach on the concept of heat distribution matrix to handle the mutual influence of the servers, in heterogeneous environments, which is novel in this context. We propose a heuristic to solve the server placement problem and we design a generic greedy framework for the online scheduling problem. We derive several single-objective heuristics (for performance, energy, cooling) and a novel fuzzy-based priority mechanism to handle their tradeoffs. Finally, we show results using extensive simulations fed with actual measurements on heterogeneous servers.

1. Introduction

The last years have witnessed the development of heterogeneity in clusters and datacenters. Two main reasons have led to this situation today. The first one is due to the maintenance and evolution of the components in the datacenters: different generations of computers are commonly seen in production datacenters since the owners are not changing everything at each update. The second reason is driven by the idea that heterogeneity might be the key to achieving energy-proportional computing [5,9], especially for high-performance computing applications.

Many recent studies alert dramatically on the energy consumption of the datacenters. For instance, Koomey's report [21] claims that today's datacenters are consuming nearly 2% of the global energy, and up to half of that is spent on cooling-related activities [33]. This results generally in very poor Power Usage Effectiveness (PUE).

In this paper, we study the multi-objective resource management problem for heterogeneous datacenters. Besides the performance criterion, we also consider the energy consumption of the servers and their thermal impact on the datacenter cooling. The aim of our work is to optimize these objectives and to explore their tradeoffs. In particular, the energy consumption is partly due to the cooling efficiency in the datacenter [25,38], which is related to both the physical placement of the servers and the scheduling strategies when jobs dynamically enter and leave the system. The latter also affects the performance and the energy consumed by the servers.

Server placement in a computer room has been relatively less studied, especially its impact on the cooling efficiency. The reason for this lack of attention is mainly due to the fact that, when servers are homogeneous, their relative positions have no impact on the performance and computing energy. However, server placement can have an impact on the cooling infrastructure. The main observation is that one server might contribute to the temperature raise at the inlets of the other servers, due to the recirculation of heat in a datacenter. Such mutual influence can be modeled by a heat distribution matrix among the servers. If one wants to keep the inlet temperature under a given threshold, the supplied

* Corresponding author.
E-mail addresses: sun@irit.fr (H. Sun), stolf@irit.fr (P. Stolf), pierson@irit.fr (J.-M. Pierson), dacosta@irit.fr (G. Da Costa).

http://dx.doi.org/10.1016/j.suscom.2014.08.005

air temperature has to be decreased accordingly by the cooling system, which in turn increases its energy consumption. In the presence of heterogeneous servers with different power consumptions and hence heat dissipation, the problem of find the optimal placement becomes complicated and, to the best of our knowledge, has not been studied. Since it is not feasible to change dynamically the positions of the servers in a datacenter, we focus on static placement to minimize the cooling cost induced by different configurations.

With a given server placement, the traditional problem of job scheduling in the heterogeneous environment remains. Many previous work (e.g., [4,40]) considered only the performance criterion and hence focused on the jobs' execution times. In order to address the power consumption issue in datacenters, however, application scheduling must employ a multi-objective approach by considering performance, energy and cooling together. To account for the fact that a scheduler has no future knowledge (jobs arrive over time), we need an online scheduling strategy. Instead of designing different independent algorithms, we design a greedy online scheduling framework that can be adapted easily by redefining the cost function, from a single objective to two or more objectives. To tackle the energy-performance tradeoff, we further introduce a fuzzy-based priority approach, which allows to explore the potential improvement in one objective while relaxing the other objective up to an acceptable range. This approach can be extended to incorporate more than two objectives in the framework. Its principle is not limited to the case at hand and can potentially be applied to other multi-objective optimization problems.

The main contributions of this paper are the following:

- A static server placement heuristic to reduce the cooling cost for the servers in a datacenter.
- A greedy scheduling framework and several cost functions to tackle single-objective scheduling (for performance, energy, and cooling).
- A fuzzy-based priority approach to handle the tradeoff between two conflicting objectives, and its extension to multi-objective optimization.

These proposals are supported by extensive simulations conducted using real hardware specifications and software benchmarks, as well as experimentally verified cooling model and heat distribution matrix [39,38]. Specifically for the hardware, a server system with high packing density and integrated cooling support is chosen for the experiments, which we believe represents well an emerging class of highly integrated energy-efficient servers. The results demonstrate the flexibility of our scheduling framework and confirm the effectiveness of the fuzzy-based approach for exploring the energy-performance tradeoff in heterogeneous datacenter environments. Our static server placement heuristic is also shown to provide much improved thermal distribution leading to significant reduction in cooling cost.

The rest of this paper is organized as follows. Section 2 formally states the system model and the scheduling problem. Section 3 describes our greedy server placement heuristic. Section 4 presents the job scheduling framework, various cost functions and the fuzzy-based priority approach. The simulation results are shown in Section 5. Section 6 reviews some related work, and Section 7 summarizes the paper and addresses future directions.

2. Problem statement

2.1. System model

Motivated by the placement of physical servers and the scheduling of high-performance computing (HPC) applications in heterogeneous datacenters, we consider the following system model: A set $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$ of m servers (or machines) needs to be placed inside a computer room (or datacenter) with a set of m rack slots, denoted by $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$.[1] Each server $M_j \in \mathcal{M}$ consists of L_j processors of the same type (possibly on different boards), but the type and the number of processors may vary for different servers, rendering the system heterogeneous. Each server consumes a *base power* U_j^{base} to support the basic operations of the infrastructure backbone, such as monitoring, networking and cooling (for instance fans). A set $\mathcal{J} = \{J_1, J_2, \ldots, J_n\}$ of n jobs arrive at the system over time, and they need to be assigned in an online manner to the servers. Each job $J_i \in \mathcal{J}$ has a release time r_i and a processor requirement l_i that must be granted in order to run on any server. To execute job J_i on server M_j incurs a processing time $P_{i,j}$ and a power consumption $U_{i,j}$, both of which are server-dependent and become known upon the job's arrival by prior profiling of the applications. In particular, the profiled application power consumption is assumed to include the leakage power.

2.2. Scheduling model

We study two orthogonal problems that deal with the placements of hardware and software, respectively. We call the two problems *static server placement* and *online job scheduling*. The former concerns the positioning of physical servers in the datacenter, which as explained in Section 3 will have an impact on the cooling energy in heterogeneous environment. The latter concerns the dynamic assignment of workloads to the servers, which will impact energy (due to both computing and cooling) as well as performance.

For the first problem of static server placement, each server needs to be physically and statically placed in advance to one of the available rack slots in the datacenter. In particular, we are looking for a mapping $\sigma : \{1, 2, \ldots, m\} \rightarrow \{1, 2, \ldots, m\}$ from rack slots to servers so that each slot S_k is filled with a server $M_{\sigma(k)}$. The objective is to minimize the cooling cost. More details about this problem will be described in Section 3.

Given a particular server placement, an online scheduling strategy is then required to assign the jobs to the servers for execution. Specifically, each arrived job $J_i \in \mathcal{J}$ must be assigned irrevocably to a server with at least l_i idle processors, and without any knowledge of the future arriving jobs. Once the job has been assigned, no preemption or migration is allowed, which is typically assumed for HPC applications since they tend to incur a significant cost in terms of data reallocation.

At any time t, the total computing power of server M_j is the sum of its base power and the power consumed for executing all jobs assigned to it, i.e.,

$$U_j^{comp}(t) = U_j^{base} + \sum_{i=1}^{n} \delta_{i,j}(t) \cdot U_{i,j} \quad (1)$$

[1] In this paper, we assume that the number of rack slots is equal to the number of servers to be placed, which represents a common scenario in small- and medium-size datacenters.