Future Generation Computer Systems 78 (2018) 302–315

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Design and implementation of energy-aware application-specific CPU frequency governors for the heterogeneous distributed computing systems[☆]

Michał P. Karpowicz*, Piotr Arabas, Ewa Niewiadomska-Szynkiewicz

NASK Research Institute, Wawozowa 18, 02-796 Warsaw, Poland
Institute of Control and Computation Engineering, Nowowiejska 15/19, 00-665 Warsaw, Poland

HIGHLIGHTS

- Design of application-specific energy-aware CPU controller is presented.
- Application-specific CPU controllers may outperform standard Linux CPU governors.
- Benchmarking methodology is proposed to identify models of CPU workload dynamics.
- Server power consumption estimate based on MSR-based measurements is proposed.

ARTICLE INFO

Article history:
Received 17 December 2015
Received in revised form
10 March 2016
Accepted 5 May 2016
Available online 22 May 2016

Keywords:
Green computing
DVFS
Data centers
Optimal control
System identification
Linux

ABSTRACT

This paper deals with the design of application-specific energy-aware CPU frequency scaling mechanisms. The proposed customized CPU controllers may optimize performance of data centers in which diverse tasks are allocated to servers with different characteristics. First, it is demonstrated that server power usage can be accurately estimated based on the measurements of CPU power consumption read from the model specific registers (MSRs). Next, a benchmarking methodology derived from the RFC2544 specification is proposed that allows to identify models of CPU workload dynamics. Finally, it is demonstrated how the identified models can be applied in the design of customized energy-aware controllers that dynamically adjust CPU frequency to the application-specific workload patterns. According to the results of experimental studies the customized controllers may outperform standard general-purpose governors of the Linux kernel both in terms of reachable server performance and power saving capabilities.

1. Introduction

To meet the demand for data processing a variety of computing resources have been combined together in large-scale heterogeneous distributed systems. This trend in technology development gives rise to a broad spectrum of computer engineering challenges, involving security, scalability, mobility, fault tolerance, energy-efficiency and performance. The design of efficient resource allocation and task scheduling algorithms has been at the cutting-edge of the related research efforts [1,2].

As the support of cloud services and HPC applications in data centers requires an ever-increasing amount of electrical power, energy-efficiency has been an important factor shaping the growth of new data processing technologies. More precisely, annual cost of energy consumed by data centers was estimated to $27 billions in 2011 (www.idc.com). According to recent estimates, the total electricity consumption accounted to ICT infrastructures increased from about 3.9% in 2007 to 4.6% in 2012. The demanded computing efficiency levels of 50 GFLOPS/W are required to be reached within the economically reasonable power consumption levels, currently ranging from 20 MW to 40 MW [3–5]. These limits are correlated with the costs of electricity and power provisioning. It is indeed commonly accepted that the problem of power consumption needs to be properly addressed, if the growth rate of ICT is to be sustained [6–9].

A straightforward way to reduce power consumption in data centers is to take advantage of the hardware-specific power saving

---

[☆] This research was partially supported by the National Science Centre (NCN) under the Grant No. 2015/17/B/ST6/01885.

* Corresponding author at: NASK Research Institute, Wawozowa 18, 02-796 Warsaw, Poland.
E-mail address: michal.karpowicz@nask.pl (M.P. Karpowicz).

http://dx.doi.org/10.1016/j.future.2016.05.011
0167-739X/

capabilities and to adjust the operating states of the computing elements to their variable workload [10–12]. These are typically exposed to the operating systems through the ACPI compatible interfaces (www.acpi.info). Power consumption of the processor (CPU) can be controlled with dynamic voltage and frequency scaling (DVFS) mechanisms capable of switching the rate of instruction execution (ACPI P-states) and the power saving levels (ACPI C-states) during idle periods [13]. In the Linux operating system these mechanisms have been provided by two kernel modules (www.kernel.org). The CPU frequency scaling process is controlled by the governors of the `cpufreq` module [14,13,15]. Behavior of the CPU during the idle periods is independently and simultaneously determined by the `cpuidle` module [16]. The modules also provide dedicated drivers, such as *intel_pstate*, designed for specific CPU architectures.

Each governor of the `cpufreq` module implements a frequency switching policy. There are several standard build-in ACPI-based governors available in the recent versions of the Linux kernel. The governors named `performance` and `powersave` keep the CPU at the highest and the lowest processing frequency, respectively. The `userspace` governor permits user-space control of the CPU frequency. Finally, the default energy-aware governor, named `ondemand`, dynamically adjusts the frequency to the observed variations of the CPU workload.

This paper deals with the problem of CPU frequency control policy design and implementation in the Linux-based servers supporting cloud services and HPC applications in heterogeneous computing systems. The context of this research, as well as the motivation for the proposed approach, is presented in Section 2. In order to illustrate the role that the CPU governors play in the resource allocation and job scheduling process, a general structure of data center control system is discussed. It is next proposed that efficiency of data processing could be improved if application-specific CPU controllers were submitted to the computing nodes along with the scheduled batch of jobs. This way both the energy consumption and the performance of the cluster nodes could be adjusted to the expected or forecasted workload patterns. The related problems of workload identification and controller design are discussed in the following sections. The problem of power consumption modeling is addressed in Section 3. Precisely, it is demonstrated how the power consumption profiles of different types of applications can be derived from the high-resolution power usage measurements read from the CPU model specific registers (MSRs). In Section 4 the problem of CPU workload dynamics identification is addressed. In particular, a dedicated benchmarking methodology is proposed together with the design of the required high-resolution probes. Finally, in Section 5 it is shown how the developed models of power consumption and processing dynamics may support the design of dedicated energy-aware controllers efficiently adjusting CPU frequency to the workload generated by a specified class of applications.

Overview of results

It is known that power consumption of a server depends on the workload generated by performed operations. Since these operations use computing resources required by the executed applications, it may be possible to design customized controllers that optimize energy-consumption and performance of a server under the application-specific workload profiles. This paper shows how the above concept can be realized and how the related engineering problems can be addressed in the environment of the Linux kernel. The addressed problems involve power consumption modeling, identification of CPU workload dynamics, controller design and implementation. The presented conclusions are supported by the results of experimental studies. Implementation details of the developed software are given as well.

First, it is demonstrated that in absence of external power-meters operating at high sampling-rate the total power usage of a server can be accurately estimated based on the high-resolution measurements read from the processor's MSRs. Experimentally identified polynomial models of low order are presented that describe power consumption profiles of different types of applications.

Second, in order to identify models of CPU workload dynamics, a general-purpose benchmarking methodology is proposed based on the RFC2544 specification. The proposed methodology identifies packet filtering dynamics of the `libpcap` module in the Linux kernel. It is experimentally verified that accurate models of data processing dynamics can be estimated this way based on the observations collected from the customized kernel probes. A conjecture is also made according to which appropriately designed benchmarking excitation signals, characterized by a specified spectral profile, could be used to identify application-specific models of data processing.

Third, it is demonstrated how the identified models can be applied in the design of customized energy-aware controllers that dynamically adjust CPU frequency (ACPI P-state) to the application-specific workload patterns. For this goal, a stochastic control problem is formulated and solved numerically. The optimal solution to the problem is a control policy that minimizes the long-run average cost of data processing operations. The policy has the form of application-specific frequency switching table parameterized by the server power consumption profile and the workload model introduced in the control problem formulation. Implementation of the obtained switching table in the form of the `cpufreq` governor is next discussed. Performance of the governor was tested in a series of webserver benchmarks and compared to the performance of standard governors provided by the Linux kernel. The obtained results show that in comparison to the `ondemand` governor the designed controller allows to reduce power consumption of the server while improving its performance or keeping it at the similar level. In particular, it is demonstrated that the customized application-specific CPU controller, tailored to heterogeneous computing environments, may outperform the general-purpose controllers, both in terms of performance and power usage efficiency.

Related work

There has been a large volume of research in the area of energy-efficient control in data centers and networks. Several recent projects may be mentioned. The ECONET project (www.econet-project.eu), co-developed by the authors of this paper, introduced dynamic power and performance control technologies, based on standby and performance scaling capabilities, improving energy-efficiency of wired network devices. Integration of the activities of major European players in networking focused on the design of energy-efficient, scalable and sustainable future networks was facilitated within the TREND project (www.fp7-trend.eu). The GAMES project (www.green-datacenters.eu) considered innovative methodologies, metrics, Open Source ICT services and tools for the active management of energy efficiency of IT Service Centres. The DEEP project (www.deep-project.eu), an innovative European response to the Exascale challenge, was focused on a novel supercomputing architecture with a matching software stack and a set of optimized grand-challenge simulation applications. The DEEP-ER project (www.deep-er.eu) extended the architecture of the DEEP project by a highly scalable I/O system. Furthermore, the project introduced new memory technology to provide increased performance and power efficiency. The CRESTA project (www.cresta-project.eu) explored how the Exascale challenge can be met by building and exploring appropriate system software for Exascale platforms. The Mont-Blanc project (www.montblanc-project.eu)