Chapter 3

Architecture and Mechanisms of Energy Auto-Tuning

Sebastian Götz
Technische Universität Dresden, Germany

Claas Wilke
Technische Universität Dresden, Germany

Sebastian Cech
Technische Universität Dresden, Germany

Uwe Aßmann
Technische Universität Dresden, Germany

ABSTRACT

Energy efficiency of IT infrastructures has been a well-discussed research topic for several decades. The resulting approaches include hardware optimizations, resource management in operating systems, network protocols, and many more. The approach the authors present in this chapter is a self-optimization technique for IT infrastructures, which takes hard- and software components as well as users of software applications into account. It is able to ensure minimal energy consumption for a user request along with a set of non-functional requirements (e.g., the refresh rate of a data extraction tool). To optimize the ratio between utility of end users and the cost in terms of energy consumption, the system needs inherent variability leading to differentiated energy profiles and mechanisms to reconfigure the system at runtime. The authors present their approach called Energy Auto-Tuning (EAT) comprised of these mechanisms and an architecture which automatically tunes the energy efficiency of IT systems.

1. INTRODUCTION

Today, developing energy efficient software has become one of the major challenges for software engineering. According to several studies, the carbon dioxide emissions caused by Information and Communication Technology (ICT) were estimated as two percent of world-wide CO2 emissions in 2007 and further increases are expected (Gartner, Inc., 2007; Climate Group, 2008). For ICT hardware, like CPUs, hard drives and network devices, several energy optimization techniques have been explored (Hewlett-Packard;
In this chapter, we propose Energy Auto-Tuning (EAT) as a solution for energy-optimizing software systems. EAT bases on Component-Based Software Development (CBSD) and takes both software and hardware components into account. Software components can provide different qualities (e.g., different refresh rates or resolutions) and can be deployed on different hardware devices (e.g., different servers). At runtime, the EAT system decides where to deploy software components once a user requests a provided service. Furthermore, the EAT system is capable to select the best of multiple existing implementations (variants of components) providing different quality of services and consuming different amounts of energy. The optimal variant is selected and deployed at the optimal server, its energy consumption.

The basic principle of Auto-Tuning (AT) can be described as a control loop. An AT system, running on a hardware infrastructure, monitors and controls itself. For each user request, the system reflects if it is able to provide the highest possible user utility for the least possible cost in terms of efforts like energy consumption in its current configuration. A system configuration in our terms is a set of software component instances deployed on component-containers, which run on servers (or, more general, computing entities). To improve efficiency, the AT system does not only need to know its current configuration, but needs to be able to compute other (probably more optimal) configurations and to reconfigure itself at runtime. Notably, reconfiguration is an activity of the system, which implies costs by itself. Hence, the effort imposed by reconfiguration needs to be taken into account additionally. The basic steps of an AT system are depicted in Figure 1.

They are: collect information about the system in its current configuration, analyze, if the system is able to optimally serve the user request, decide for a reconfiguration if another, more efficient configuration exists, and act, i.e. perform the actual reconfiguration. The general objective function of the overall system is described by Formula 1:

$$\text{efficiency} = \frac{\sum_{i=1}^{n} (\text{utility}_i \cdot w_i^u)}{\sum_{j=1}^{m} (\text{effort}_j \cdot w_j^e)}$$

Efficiency is defined as the ratio of the sum of weighted utilities and the sum of weighted efforts to achieve these utilities. The number of utilities is denoted by $n$, whereas the number of efforts is denoted by $m$. The variables $w_i^u$ and $w_j^e$ (i.e., weights) are used to prioritize certain utilities and efforts respectively. Energy efficiency is a special case of this formula, where the only effort of interest is the energy consumption (cf. Formula 2).