UMEÅ UNIVERSITET

Energy-efficient Cloud Computing:

Autonomic Resource Provisioning for Datacenters

Selome Kostentinos Tesfatsion
ሰሎሜ ቆስጠንጢኖስ ተስፋጽዮን

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Salens namn eller beteckning, byggnad MA121, MIT-huset, Måndag 16:e dagen den April månad, 2018, kl. 10:15.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Dr Laurent LEFEVRE, Permanent Researcher, INRIA Lyon Université de Lyon, Lyon, France.

Department of Computing Science

Organization
Umeå University
Computing Science

Document type
Doctoral thesis

Date of publication
23rd March 2018

Author
Selome Kostentinos Tesfatsion

Title
Energy-efficient Cloud Computing: Autonomic Resource Provisioning for Datacenters.

Abstract

Energy efficiency has become an increasingly important concern in data centers because of issues associated with energy consumption, such as capital costs, operating expenses, and environmental impact. While energy loss due to suboptimal use of facilities and non-IT equipment has largely been reduced through the use of best-practice technologies, addressing energy wastage in IT equipment still requires the design and implementation of energy-aware resource management systems. This thesis focuses on the development of resource allocation methods to improve energy efficiency in data centers. The thesis employs three approaches to improve efficiency for optimized power and performance: scaling virtual machine (VM) and server processing capabilities to reduce energy consumption; improving resource usage through workload consolidation; and exploiting resource heterogeneity.

To achieve these goals, the first part of the thesis proposes models, algorithms, and techniques that reduce energy usage through the use of VM scaling, VM sizing for CPU and memory, CPU frequency adaptation, as well as hardware power capping for server-level resource allocation. The proposed online performance and power models capture system behavior while adapting to changes in the underlying infrastructure. Based on these models, the thesis proposes controllers that dynamically determine power-efficient resource allocations while minimizing performance penalty.

These methods are then extended to support resource overbooking and workload consolidation to improve resource utilization and energy efficiency across the cluster or data center. In order to cater for different performance requirements among collocated applications, such as latency-sensitive services and batch jobs, the controllers apply service differentiation among prioritized VMs and performance isolation techniques, including CPU pinning, quota enforcement, and online resource tuning.

This thesis also considers resource heterogeneity and proposes heterogeneous-aware scheduling techniques to improve energy efficiency by integrating hardware accelerators (in this case FPGAs) and exploiting differences in energy footprint of different servers. In addition, the thesis provides a comprehensive study of the overheads associated with a number of virtualization platforms in order to understand the trade-offs provided by the latest technological advances and to make the best resource allocation decisions accordingly. The proposed methods in this thesis are evaluated by implementing prototypes on real testbeds and conducting experiments using real workload data taken from production systems and synthetic workload data that we generated. Our evaluation results demonstrate that the proposed approaches provide improved energy management of resources in virtualized datacenters.

Keywords
cloud computing, datacenter, energy efficiency, performance management, virtualization

Language	ISBN	ISSN	Number of pages
English	978-91-7601-862-0	0348-0542	63 + 6 papers