WiP Abstract: Intelligent Power- and Performance-aware Tradeoffs for Multicore Servers in Cloud Data Centers *

Faruk Caglar, Shashank Shekhar, Kyoungho An, and Aniruddha Gokhale
Dept of EECS, Vanderbilt University, Nashville, TN 37235, USA
{faruk.caglar, shashank.shekhar, kyoungho.an, a.gokhale}@vanderbilt.edu

ABSTRACT

Cloud service providers are seeking new solutions for reducing the energy consumption in their data centers. Our work in developing middleware that is integrated with a Virtual Machine Manager (VMM) to lower the energy consumption while maximizing the performance of or minimizing the service level agreement (SLA) violations for applications. To that end our solution leverages techniques such as Dynamic voltage/frequency scaling (DVFS) and server consolidation through live migration of virtual machines from an overloaded host machine to a less-loaded one.

Categories and Subject Descriptors

D.4.8 [**Operating Systems**]: Performance—*Energy*

General Terms

Performance

Keywords

energy, middleware, neural networks, multicore

1. ONGOING WORK IN POWER AND PERFORMANCE TRADEOFFS IN DATA CENTERS

Energy conservation in data centers is increasingly a key consideration for cloud data center operators to keep both costs low as well as alleviate the adverse impact on environment [1]. In doing so they must continue to meet application performance requirements. To address these issues, in this paper, we present preliminary ideas on how to reduce the power consumed by the host machines while meeting the service level agreements (SLAs) of the system in virtualized data centers. Our system provides a novel middleware solution that integrates dynamic voltage/frequency scaling (DVFS) and an intelligent power and performance-aware server consolidation algorithm within the context of a virtual machine manager (VMM).

Two approaches are employed by our middleware, which exploits the multicore server architecture and comprises the following:

• A mechanism integrated with the hook manager of the VMM for dynamically scaling the frequency of CPU cores of host machines by communicating with the VMM. If a core is not being used by the system, the frequency of that core is decreased and the reverse strategy is followed if the core will be in use.

• An intelligent power- and performance-aware server consolidation strategy which ensures reduction in energy consumption while satisfying the performance/SLA requirements in virtualized data centers. Resource information of a host machine is used in deciding where to place a VM from an over-loaded host machine to a less-loaded one. A two layer artificial neural network (ANN) is employed at the core of the server consolidator algorithm to make a decision on when to migrate functionality and to where.

The ANN uses CPU utilization, memory utilization, network input bytes, network output bytes, number of cores, and frequency of the cores as inputs. The estimated power consumption and SLA violations are the system outputs of the ANN. Our current work is designing and evaluating the efficacy of our solution in meeting the dual objectives outlined above.

*This work was supported in part by AFRL C2ORES, and NSF Awards CNS 0915976 and CAREER CNS 0845789. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Airforce Research Lab or NSF.

ICCPS '13 Philadelphia, PA, USA
ACM 978-1-4503-1996-6/13/04.

2. REFERENCES

[1] US Environmental Protection Agency, "Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431," www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf, 2007.