Energy-aware and Machine Learning-based Resource Provisioning of In-Memory Analytics on Cloud

Hosein Mohammadi Makrani, Hossein Sayadi, Devang Motwani, Han Wang,
Setareh Rafatirad, and Houman Homayoun
George Mason University

ABSTRACT

In this work, we propose a proactive online resource provisioning methodology that addresses the challenge of resource provisioning for IMC workloads in heterogeneous cloud platforms consist of diverse types of servers. As cloud platforms provide a wide range of server configuration choices [4], and the applications' performance and power consumption changes at run-time [3] and depends on the chosen configuration, resource provisioning in cloud platforms is a challenging optimization problem with a large search space to navigate. Our methodology proactively assigns a suitable hardware configuration to IMC program for energy-efficiency (EDP) optimization at run-time before any significant change occurs in application's behavior. This helps to save energy without sacrificing performance [2, 7]. We address these challenges by first characterizing diverse types of IMC workloads across different types of server architectures. The characterization aids to accurately capture applications' behavior [1] and train machine learning models [5, 6]. We use time series neural network to predict the next phase of an application. Our approach then uses artificial neural networks to estimate the performance and power consumption of predicted phase of application on various server configurations. Further, we use the genetic algorithm to distinguish close-to-optimal configuration to minimize EDP. Compared to Oracle scheduler, our methodology achieves 93% accuracy to allocate the right resource for each phase of the program. Our methodology improves the performance by 21% and the EDP by 40% on average, compared to the default scheduler.

CCS CONCEPTS

• **Computer systems organization → Cloud computing**;

KEYWORDS

Resource provisioning, cloud, energy-aware, machine learning

ACM Reference Format:
Hosein Mohammadi Makrani, Hossein Sayadi, Devang Motwani, Han Wang, Setareh Rafatirad, and Houman Homayoun. 2018. Energy-aware and Machine Learning-based Resource Provisioning of In-Memory Analytics on Cloud. In *Proceedings of ACM Symposium on Cloud Computing , Carlsbad, CA, USA, October 11–13, 2018 (SoCC '18),* 1 pages. https://doi.org/10.1145/3267809.3275459

1 METHODOLOGY

The proposed methodology is a configuration tuning methodology that automatically adjusts the hardware configuration assigned to a Virtual Machine (VM) in a proactive manner in order to dynamically optimize the energy efficiency of a given IMC program on a given heterogeneous cluster of servers. Our methodology consists of four major components that all of them running on one server as a manager. Components are as follow: predictor, estimator, explorer, and decision maker.

Predictor: The first component of our methodology is the predictor that predicts the next phase of application and its microarchitectural signature based on the current and previous states. For this purpose, we employ time series neural networks.

Estimator: Given the predicted signature and corresponding server configuration, we use an estimator to estimate the application's energy consumption and performance in term of energy-delay product (EDP) metric. In order to model the EDP of each application on different platforms, in this work we used Artificial Neural Networks (ANN).

Explorer: Searching component automatically searches for the best platform and configuration that minimizes the EDP for a given IMC application. As minimizing the EDP of an IMC program in the global space of configuration parameters is a complex large space to explore with many local optima, genetic algorithm (GA) is employed. GA uses techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover.

Decision Maker: After finding the optimum configuration, decision maker divides the resource allocation problem into smaller sub-problems using a hierarchical approach. The first escalation level works locally on a host and attempts to change the host resources, for example the amount of storage or memory that is allocated to the VM. If there is no appropriate VM available in level 1, the second escalation level is called where the manager creates a new VM on an appropriate node or migrates the VM to a node that has enough available resources.

REFERENCES

[1] H. M. Makrani and et al. Memory requirements of Hadoop, Spark, and MPI based big data applications on commodity server class architectures. In *IISWC'17.*

[2] H. M. Makrani and et al. Understanding the role of memory subsystem on performance and energy-efficiency of Hadoop applications. In *IGSC'17.*

[3] H. M. Makrani and et al. 2018. A comprehensive Memory Analysis of Data Intensive Workloads on Server Class Architecture. In *MEMSYS'18.*

[4] Hosein Mohammadi Makrani and Houman Homayoun. MeNa: A memory navigator for modern hardware in a scale-out environment. In *IISWC'17.*

[5] Hossein Sayadi and et al. Customized machine learning-based hardware-assisted malware detection in embedded devices. In *IEEE TrustCom-18.*

[6] Hossein Sayadi and et al. Ensemble learning for effective run-time hardware-based malware detection: A comprehensive analysis and classification. In *DAC'18.*

[7] Hossein Sayadi and et al. Machine learning-based approaches for energy-efficiency prediction and scheduling in composite cores architectures. In *ICCD'17.*

SoCC '18, October 11–13, 2018, Carlsbad, CA, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6011-1/18/10.
https://doi.org/10.1145/3267809.3275459