Accepted Manuscript

Configuring in-memory cluster computing using random forest

Zhendong Bei, Zhibin Yu, Ni Luo, Chuntao Jiang, Chengzhong Xu, Shengzhong Feng

PII: S0167-739X(16)30516-7
DOI: http://dx.doi.org/10.1016/j.future.2017.08.011
Reference: FUTURE 3606

To appear in: *Future Generation Computer Systems*

Received date : 5 November 2016
Revised date : 9 April 2017
Accepted date : 8 August 2017

Please cite this article as: Z. Bei, Z. Yu, N. Luo, C. Jiang, C. Xu, S. Feng, Configuring in-memory cluster computing using random forest, *Future Generation Computer Systems* (2017), http://dx.doi.org/10.1016/j.future.2017.08.011

Configuring In-memory Cluster Computing Using Random Forest

Zhendong Bei[a,b], Zhibin Yu[a,*], Ni Luo[a,b], Chuntao Jiang[a], Huiling Zhang[a], Chengzhong Xu[a,c], Shengzhong Feng[a]

[a]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, P.R.China.*
[b]*Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, P.R.China.*
[c]*Department of Electrical & Computer Engineering Wayne State University, Detroit, Michigan.*

Abstract

Recently, in-memory cluster computing (IMC) gains momentum because it accelerates traditional on-disk cluster computing (ODC) up to several tens of times for iterative and interaction applications. The most popular IMC framework is Spark and it has more than 100 configuration parameters. However, it is unclear how significantly these parameters affect the system performance because IMC is a quite new computing paradigm. Consequently, there is yet no study addressing how to optimally configure IMC frameworks.

In this paper, we first investigate how significantly the configuration parameters affect the performance of Spark workloads. We find that the configuration caused performance variation can be as large as 20.7, indicating configuring Spark workloads is extremely important to their performance. However, manually configuring Spark workloads is notoriously difficult because there are so many configuration parameters which might interfere with each other in a complex way. To address this issue, we propose an approach to Automatically Configure Spark workloads, named ACS. It firstly constructs performance models as functions of Spark configuration parameters by using random forest which is an ensemble learning algorithm. Subsequently, ACS leverages genetic algorithm to

*Corresponding author

Email addresses: zd.bei@siat.ac.cn (Zhendong Bei), zb.yu@siat.ac.cn (Zhibin Yu), ni.luo@siat.ac.cn (Ni Luo), ct.jiang@siat.ac.cn (Chuntao Jiang), hl.zhang@siat.ac.cn (Huiling Zhang), cz.xu@siat.ac.cn (Chengzhong Xu), sz.feng@siat.ac.cn (Shengzhong Feng)