Information Sciences 433–434 (2018) 401–416

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

MLBox: Machine learning box for asymptotic scheduling

Mihaela-Andreea VASILE, Florin POP*, Mihaela-Cătălina NIŢĂ, Valentin CRISTEA

Faculty of Automatic Control and Computers, University Politehnica *of Bucharest, Romania*

ARTICLE INFO

Article history:
Received 15 March 2016
Revised 21 November 2016
Accepted 2 January 2017
Available online 3 January 2017

Keywords:
Asymptotic scheduling
BoT scheduling
DAG scheduling
Scheduling heuristics
Machine learning
Datacenters

ABSTRACT

As the usability of Cloud-based solutions has increased for various types of users with different needs, from scientists that want to process big data sets collected from sensors or business analysts that want to take decisions based on the huge amount of gathered data to simple users that store or share documents via a Cloud platform, the generated data is increasing more and more. For example, the ATLAS and other detectors at CERN generate petabytes of data and Facebook stores data with a rate of around 600 TB daily. In the current context, efficient scheduling for Big Data applications is a challenge and an appropriate scheduling technique is required for different types of incoming requests. In this paper we propose a scheduling algorithm for different types of computation requests: independent tasks, like bag of tasks (BoT) model or tasks with dependencies modeled as directed acyclic graphs (DAG), and they will be scheduled for execution in a Cloud datacenter. The tasks in the requests are scheduled on the available resources using the suitable scheduling algorithm for each request. We rely on a machine learning toolbox, named as MLBox, to find what algorithm should be used for a certain request. We implemented four heuristics for scheduling BoTs and four heuristics for DAGs scheduling and generated the training data for the machine learning algorithm by running multiple traditional scheduling algorithms and selecting the 'best' one for a given request. We evaluate the performance by comparing the scheduling of different tasks requests using some of the traditional algorithms and our machine learning based scheduling algorithm.

1. Introduction

The scheduling problem refers to assigning tasks on resources in an almost optimal manner. This is one of the key topics in the current context of distributed systems as the schedulers, part of Cloud platforms (like Apollo [4], Bistro [14], Paragon [10], DejaVu [38], etc.), have to face new challenges caused specially by Big Data applications like processing huge amounts of heterogeneous data, generated in a short period of time, via data flows or on-line streaming, which has to be managed very fast (read, write, filter, compute statistics).

Cloud solutions enable users the access via Internet to various types of resources such as existing applications in the Cloud, frameworks that can be used for development of custom built applications (Hadoop, Spark, Flink, RabbitMQ, etc.), access to Virtual Machines (VMs) for installing operating systems and also storage and sharing solutions. Therefore, the Cloud is now a significant choice for multiple types of users, be them common individuals, scientists or technical users. The

* Corresponding author.
E-mail addresses: mihaela.a.vasile@gmail.com (M.-A. VASILE), florin.pop@cs.pub.ro (F. POP), catalina.nita23@gmail.com (M.-C. NIŢĂ), valentin.cristea@cs.pub.ro (V. CRISTEA).

http://dx.doi.org/10.1016/j.ins.2017.01.005
0020-0255/