Future Generation Computer Systems 82 (2018) 1–11

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A survey and taxonomy of resource optimisation for executing bag-of-task applications on public clouds

Long Thai [a], Blesson Varghese [b], Adam Barker [a],*

[a] School of Computer Science, University of St Andrews, Fife, UK
[b] School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, Belfast, United Kingdom

HIGHLIGHTS

- Bag-of-Task (BoT) applications execute on heterogeneous cloud settings.
- User-defined constraints and objectives are measure of quality of service.
- Scheduling plans are generated using exact and heuristic algorithms.
- Future research will need to focus on awareness of cross cloud scheduling.
- Performance estimation of BoT will be required for optimising cloud resources.

ARTICLE INFO

Article history:
Received 28 March 2017
Received in revised form 5 October 2017
Accepted 23 November 2017
Available online 9 December 2017

Keywords:
Cloud computing
Bag-of-Task
Execution optimisation
Resource usage optimisation

ABSTRACT

Cloud computing has been widely adopted due to the flexibility in resource provisioning and on-demand pricing models. Entire clusters of Virtual Machines (VMs) can be dynamically provisioned to meet the computational demands of users. However, from a user's perspective, it is still challenging to utilise cloud resources efficiently. This is because an overwhelmingly wide variety of resource types with different prices and significant performance variations are available.

This paper presents a survey and taxonomy of existing research in optimising the execution of Bag-of-Task applications on cloud resources. A BoT application consists of multiple independent tasks, each of which can be executed by a VM in any order; these applications are widely used by both the scientific communities and commercial organisations. The objectives of this survey are as follows: (i) to provide the reader with a concise understanding of existing research on optimising the execution of BoT applications on the cloud, (ii) to define a taxonomy that categorises current frameworks to compare and contrast them, and (iii) to present current trends and future research directions in the area.

1. Introduction

Cloud computing has become a sizeable industry and allows users, including industry and academic organisations to rent resources. According to a Business Insider report, the revenues of Amazon Web Services (AWS) and Microsoft Azure have exceeded $6 billion a year [1]. The adoption rate of cloud computing is high, according to a report by RightScale [2]; nearly 88% of 930 organisations considered in the report took advantage of cloud computing.

There are usually two parties involved in cloud computing: *cloud providers* and *cloud users*. Cloud providers build and maintain data centres, such as Amazon or Google. They manage and maintain the physical infrastructure on which the cloud is running and define the types of resources that are available to users and the associated pricing.

Cloud users require computational resources to run applications. Users range from commercial companies, academic organisations, or private users who deploy and run a few or all of their applications on the cloud. Therefore, cloud users do not need to pay attention to the deployment and maintenance of the physical infrastructure. However, they are responsible for utilising the resources offered by the providers to build their own *cloud cluster*. We define a cloud cluster as a collection of VMs that is used to achieve the intended goals of a workload deployment.

Research in cloud computing can be broadly classified on two different points of view, both of which are necessary for developing next-generation cloud computing systems [3]. The first one aims to help cloud providers efficiently build, manage and operate cloud

* Corresponding author.
E-mail addresses: long.t.thai@gmail.com (L. Thai), varghese@qub.ac.uk (B. Varghese), adam.barker@st-andrews.ac.uk (A. Barker).
URLs: http://www.blessonv.com (B. Varghese), http://www.adambarker.org/ (A. Barker).

https://doi.org/10.1016/j.future.2017.11.038
0167-739X/© 2017 Elsevier B.V. All rights reserved.

infrastructure. Research in this direction can be categorised as *cloud data centre optimisation*, in which the resources are represented as Physical Machines (PMs) that a cloud provider owns and maintains. The optimisation technique, referred to as *VM placement*, aims to map VMs onto PMs in order to minimise the number of allocated PMs.

The second category is based on *cloud usage optimisation*, which takes the user's point of view into account and deals with optimisation tasks such as: how does a user make a decision about which resources to utilise, or when to scale an application on the cloud. Inputs describing the cloud environment, a user's application(s) and requirements are taken into account to determine a course of action, such as resizing a cloud cluster and distributing workloads among VMs. Since the physical infrastructure is abstracted away from the user, research in this direction normally assumes that resources are unlimited and focuses on minimising the incurred monetary costs associated with renting VMs.

Users run a variety of applications or workloads on the cloud, ranging from simple Web applications, workflows and frameworks which support computationally-intensive applications, such as MapReduce, and Spark. A Bag-of-Task (BoT) application is one class of workload that is commonly used on the cloud and consists of many independent tasks, each of which can be executed by any machine in any order. They can be executed concurrently by many different machines. For instance, a simulation application, e.g. Monte Carlo simulation [4], is a BoT application in which each execution represents a task. Similarly, a parameter sweep application [5] is a BoT application in which each task corresponds to one combination of parameters. However, a MapReduce [6] application is not a BoT application since the Reduce phase must wait for the Map phase to complete. The Map and Reduce phases can be considered as two different BoT applications. This survey paper focuses on the cloud usage optimisation for BoT applications.

We have selected to survey BoT on the cloud because they are widely utilised by both scientific and commercial organisations. These applications are large and too complex to be executed on a single machine. They are also the dominant applications that are submitted to and utilise CPU time in grid environments [7]. Similarly, companies, such as Facebook, report that the jobs running on their own internal data centres are mostly independent tasks [8].

Even though there have been multiple surveys regarding research in cloud optimisation, we believe that they do not provide a holistic view of the research in cloud usage optimisation. For instance, the surveys of Fakhfakh et al. [9] and Wu et al. [10] focus on workflow applications. Surveys in this area are usually based on the point of view of cloud providers [11] or treat optimising cloud usage as one of many aspects of cloud data centre management [12].

This paper is distinguished from existing surveys by focusing on the methodologies that can be used by cloud users, who do not have a complete view of the underlying infrastructure of the public cloud. In this survey, we set out to review the existing publications regarding optimising the cloud resource from a user's point of view. Furthermore, we focus only on BoT applications.

The goal of this survey is threefold: (i) to provide a holistic and concise view of the current state-of-the-art in cloud usage optimisation for BoT applications, (ii) to define a taxonomy that categorises current research to compare and contrast existing frameworks, and (iii) to present current trends and employ the taxonomy as a guide for furthering research in the area.

1.1. Data collection

The research publications used in this survey were collected in September 2017 via Google Scholar. To ensure the quality of the publication, we selected articles from high-impact journals, such as IEEE Transactions on Services Computing and IEEE Transactions on Parallel and Distributed Systems, and top-tier conference venues, such as IEEE Conference of Cloud Computing (CloudCom), IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), and IEEE International Conference on Cloud Computing (CLOUD).

We set out the following criteria for a publication to be selected for this survey:

- The application presented in the publication must be a BoT application; we did not consider publications that presented workflow or user-facing applications.
- The execution results presented in the publication must be performed fully or partly on the cloud; we did not consider other resource environments, such as grids.
- The monetary cost incurred in executing an application must be considered in the publication, which is a unique characteristic of the cloud environment.
- The assumption in the publication must be that the cloud is a black-box environment, e.g., a public cloud in which a user has little to no control over internal operations.

The above criteria were set in line with our survey goals — develop a taxonomy of resource optimisation for executing BoT applications on public clouds. At the end of the publication selection phase, there were 31 publications that satisfied the criteria and are used as the basis for this survey.

1.2. Organisation

The organisation of this survey is shown in Fig. 1. We firstly consider the current research of BoT applications in Section 2. As previously indicated we present this from the perspective of the methodologies that a cloud user can adopt rather than the techniques used in the underlying infrastructure or middleware of public clouds that is usually inaccessible to a user. Both scheduling of BoT applications on a *homogeneous* cloud and a *heterogeneous* cloud are considered. We refer to homogeneous clouds as environments that use the same type of VMs in public clouds, and to heterogeneous clouds as environments where different VM types are used. We explore scheduling in the context of hybrid clouds, spot VMs, and on-demand VMs for both homogeneous and heterogeneous clouds and in addition for reserved VMs in homogeneous clouds.

The taxonomy we propose in Section 3 is based on six themes, namely functionality, requirements, parameter estimation, dynamic scheduling, solving methods and application heterogeneity. For each of these themes, we first present an overview and then the associated review of the literature. Our survey then uses the above taxonomy for summarising four current trends that are seen in BoT scheduling, Section 4.1. We use these to chart out three future directions for optimising cloud usage for executing BoT applications in Section 4.2.

Although the structure presented in Fig. 1 is created to survey research specific to scheduling BoT applications, it may be broadly used for other applications, such as workflows or user-facing applications.

2. Current research

In this section, we survey research that focuses on developing frameworks for scheduling the execution of BoT jobs on the cloud. This survey is focused on BoT jobs on cloud resources and therefore alternate application types (such as workflows) and resource models (such as the grid or clusters) are not considered. This survey assumes a cloud user's point-of-view, which treats the cloud as a black box and the user may not have control over its internal