Simulation Modelling Practice and Theory 55 (2015) 46–62

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

A cloudification methodology for multidimensional analysis: Implementation and application to a railway power simulator

Silvina Caíno-Lores *, Alberto García Fernández, Félix García-Carballeira, Jesús Carretero Pérez

Computer Science and Engineering Department, University Carlos III of Madrid, Avda. Universidad 30, 28911 Leganes, Madrid, Spain

ARTICLE INFO

Article history:
Received 18 December 2014
Received in revised form 20 March 2015
Accepted 9 April 2015
Available online 29 April 2015

Keywords:
Railway simulator
Migration
MapReduce
Cloud computing
Many-task computing

ABSTRACT

Many scientific areas make extensive use of computer simulations to study complex real-world processes. These computations are typically very resource-intensive and present scalability issues as experiments get larger even in dedicated clusters, since these are limited by their own hardware resources. Cloud computing raises as an option to move forward into the ideal unlimited scalability by providing virtually infinite resources, yet applications must be adapted to this new paradigm. This process of converting and/or migrating an application and its data in order to make use of cloud computing is sometimes known as cloudifying the application. We propose a generalist cloudification method based in the MapReduce paradigm to migrate scientific simulations into the cloud to provide greater scalability. We analysed its viability by applying it to a real-world railway power consumption simulatior and running the resulting implementation on Hadoop YARN over Amazon EC2. Our tests show that the cloudified application is highly scalable and there is still a large margin to improve the theoretical model and its implementations, and also to extend it to a wider range of simulations. We also propose and evaluate a multidimensional analysis tool based on the cloudified application. It generates, executes and evaluates several experiments in parallel, for the same simulation kernel. The results we obtained indicate that out methodology is suitable for resource intensive simulations and multidimensional analysis, as it improves infrastructure's utilization, efficiency and scalability when running many complex experiments.

1. Introduction

Scientific simulations constitute a major set of applications that attempt to reproduce real-world phenomena in a wide range of areas such as engineering, physics, mathematics and biology. Their complexity usually yields a significant resource usage regarding CPU, memory, I/O or a combination of them.

In order to properly scale these applications, they can be distributed to a cluster or grid. While these approaches have proved successful, they often rely on heavy hardware investment and they are tightly conditioned by the available resources. This de facto limits actual scalability and the addressable simulation size. Since sharing resources across multiple clusters implies several limitations, cluster applications cannot be considered sustainable, because their scalability is strongly dependant on the cluster size.

* Corresponding author.
E-mail addresses: scaino@arcos.inf.uc3m.es (S. Caíno-Lores), agarcia@arcos.inf.uc3m.es (A.G. Fernández), fgarcia@arcos.inf.uc3m.es (F. García-Carballeira), jcarrete@arcos.inf.uc3m.es (J.C. Pérez).

http://dx.doi.org/10.1016/j.simpat.2015.04.002
1569-190X/© 2015 Elsevier B.V. All rights reserved.

Despite scientific simulations will likely benefit from the upcoming exascale infrastructures [1], the challenges that must be overcome –power consumption, processing speed and data locality, for instance [2]– will probably rise again in the future as applications become more complex. Therefore, the ideal situation of unlimited scalability seems difficult to reach with this approach.

Another option is cloud computing, which has been increasingly studied as an alternative to traditional grid and high-performance distributed environments for resource-demanding and data-intensive scientific simulations [3]. Cloud computing emerged with the idea of elasticity: virtual unlimited resources obtainable on-demand with minimal management effort [4]. It would enable the execution of large simulations with virtual hardware properly tailored to fit specific use cases like memory-bound simulations, CPU-dependant computations or data-intensive analysis. It holds further advantages, such as elasticity, automatic scalability and instance resource selectivity. Moreover, its so-called pay-as-you-go model allows to adjust the required instances to the particular test case size while cutting-down the resulting costs.

Furthermore, recent advances in cloud interoperability and cloud federations can contribute to separate application scalability from datacenter size [5,6]. From that point of view, applications would become more sustainable, as they can be operated in a more flexible way through heterogeneous hardware, cross-domain interactions, and shared infrastructures.

There are several issues that should be tackled in order to develop a sustainable application, such as:

- Virtual unlimited scalability could be achieved by reducing the number architectural bottlenecks, such as network communications or master-node dependencies. This would minimize the added overhead of working with more nodes, making a better use of the available resources.
- By making the application platform independent, we can aggregate computational resources possibly located in different places, hence local data center size would not be a limitation. Moreover, we can exploit cluster and cloud resources simultaneously following an hybrid scheme.
- An application that could behave in a flexible manner efficiently would be able to scale up or down easily according to instantaneous user needs, thus adapting computing resources to specific simulation sizes and deadlines.
- If the application already exists and has to be adapted, it is desirable to minimize the impact on the original code, thus performing the minimal modifications needed to achieve the aforementioned objectives.

Given the former, resource-intensive scientific simulations hold potential scalability issues on large cases, since standalone and cluster hardware may not satisfy simulation requirements under such stress circumstances. Therefore, in previous work we have explored the possibility of performing a paradigm shift from single-node HPC computations to a data-centric model that would distribute the simulation load across a set of virtual instances [7,8]. In this paper we propose a generic methodology to transform scientific simulations into a cloud-suitable data-centric scheme via the MapReduce framework. Moreover, in this paper we provide an optional experiment generation stage that allows users to configure a full set of simulations with a varying parameter for solution optimization purposes. This multidimensional analysis capability is translated to a many-task problem within our methodology.

The processes mentioned are illustrated by means of a real-world application, a simulator which calculates power consumption on railway installations. This simulator, starting from the train movements (train position and consumption), calculates the instantaneous power demand (taking into account all railway elements such as tracks, overhead lines, and external consumers) indicating whether the power provisioned by power stations is enough or not. Simulator internals consist on composing the electric circuit on each instant, and solving that circuit using modified nodal analysis. The starting version of the simulator, based on multi-threading, is memory bounded, strongly limited by the number of instants to be simulated simultaneously (and therefore by the number of threads). The resulting performance is evaluated on Amazon Elastic Compute Cloud running Hadoop YARN MapReduce.

The rest of this paper is organized as follows: Section 2 discusses related works, Section 3 describes our proposed methodology, Section 4 illustrates the cloudification transformation method on a particular use case, Section 5 evaluates how the resulting design implementation on Hadoop MapReduce 1.1.2 (MRv1) and Hadoop YARN Mapreduce 2.2.0 (MRv2) behaves on both a local cluster and Amazon Elastic Compute Cloud (EC2), Section 6 describes the process of transforming the methodology into a multidimensional analysis by means of a many-task experiment generation and evaluation process and, finally, Section 8 provides key ideas as conclusions and some insight in future work.

2. Related work

Scientific applications and their adaptability to new computing paradigms have been dragging increasing attention from the scientific community in the last few years. The applicability of the MapReduce scheme for scientific analysis has been notably studied, specially for data-intensive applications, resulting in an overall increased scalability for large data sets, even for tightly coupled applications [9].

Hadoop MapReduce is nowadays widely used as base platform for new programming languages and architectures. Hadoop MapReduce is used in Pig Latin [10], an associative language used in Yahoo for taking advantage of both declarative languages and map-reduce programming style. This approach is strongly focused on processing data sets, and does not tackle the issue of scientific workflows. Apache Hive [11] and Bigtable [12] are two storage systems developed on the top of